

Rays of hope in eastern Europe

Although much of Russia struggles to survive, other former Soviet countries are building new research systems with varying success. A critical need is to avoid western domination of their programmes.

Ten years can be a short time in politics, at least in politics of the dimensions that shook central and eastern Europe in the 1990s. For scientists, much changed literally overnight when communism fell. Yet, one thing was all too clear at a Dresden meeting last week of east and west European scientists, organized by *Nature* and the Volkswagen Foundation, and sponsored by NATO. This was that there is still a long way to go in replacing the systems of science that were so quickly destroyed.

"Emancipation from ideology is less easy than you'd like to think," Andrei Plesu, rector of the New Europe College in Bucharest, told the audience. The restrictive, centralized, communist system desperately needed reform, but the "traps of the cure" also need to be acknowledged, he argued.

It came as no surprise at the meeting to learn that lack of national money for research has undermined efforts to build internationally competitive science bases in all central and eastern European countries. But Plesu's "traps" are less about money shortages *per se* than, paradoxically, about the isolation that sudden decentralization brought, as a consequence of an immediate loss of all points of reference, and how this fact conspired to allow eastern science agendas to be set by the west, sometimes inappropriately. Former communist countries tended to accept any form of western sponsorship, pretending that the gift was exactly what they wanted and needed, for fear of being left empty-handed, he said.

The mood of the meeting was cautiously optimistic, the bottom line being that things are clearly better than they were. But speakers referred time and again to the unwitting agenda-setting for the east by the west, which in part results from the same international collaboration and direct-aid programmes that served to save the best science from destruction, and helped build up some islands of excellence. Combined with the introduction of competitive distribution of small national research budgets, and in the absence of strongly defined national priorities, this tended to support the status quo. Thus physics- and chemistry-based disciplines have been disproportionately supported, as these were strengths in communist times. Yet energy and environment research, so visibly necessary in the former communist countries, have failed to develop.

Funding failings

Dependence on the west during the transition could not have been avoided, and it was also no surprise to learn at the meeting that those doing best from it are the 12 European Union (EU) candidate countries: those being considered for EU membership in the next few years. These countries, which include Poland, Hungary and the Czech Republic, are now participating fully in the fifth Framework programme of research. But this participation requires matching funds which consume most of their small national research budgets, thus compounding the problem of research agendas being set in the west.

What is to be done? The meeting made clear that in the foreseeable future the answer for the non-EU candidate countries lies solely in ensuring continuation of foreign funds, particularly those supporting collaborative projects, and those, provided for example by the

Volkswagen Foundation, for supporting young scientists. For countries such as Russia, the situation for science remains desperate and simple survival is the priority. On average, between a third and a half of the research funds of top-quality Russian research institutions come from foreign sources. The Volkswagen Foundation, one of the largest supporters of science in the central and eastern European countries, is aware of the impact that the impending end of its support programme will have.

For most candidate countries, survival and even growth of science is assured, but governments must now be prepared to invest more robustly in their national programmes, to allow a reasonable level of independence to develop. These countries can now afford it.

New targets

Some speakers also pointed out that scientists themselves should be proactive in identifying emerging areas of research and positioning themselves in them. Günter Stock, head of research at the German pharmaceutical company Schering, advised scientists from central and eastern European countries to establish themselves in many of the new, and still relatively uncrowded, research areas of the post-genomic era. At least 5,000 new candidate drug targets are likely to be identified from genomic information, more than the pharmaceutical industry wants to handle itself, he said. Validating new targets sometimes requires skills that have declined in the west, such as old-fashioned whole-animal physiology, and it also requires a lot of clinical research.

It is widely appreciated that there are fewer constraints on this type of work — as well as stem-cell research which Stock also advocates — in the east than the west. Pharmaceutical companies are actively searching for such collaborations and Stock said scientists should facilitate matchmaking by advertising themselves, perhaps on the Internet. Would this mean imposing yet another agenda set by the west, this time by the pharmaceutical industry? Perhaps, but there are worse ways to survive, and it would at least allow research to move to the forefront of applied biology.

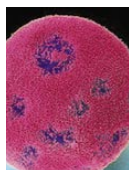
One example of how scientists can be successfully proactive is in Estonia. Being a small and therefore flexible country, Estonia was able to institute reform of its scientific institutions more efficiently and completely than most. It dissolved its Soviet-style Academy of Sciences early on and merged the academy's institutes with universities, which have been granted full autonomy. In this atmosphere, scientists have found the self-confidence to determine their own research agendas, bottom-up.

The boldest example is perhaps the creation of the Estonian Population Genetics Project, whereby the blood of every Estonian will be DNA profiled, to search for disease-associated genes. Scientists at the University of Tartu are already establishing start-up companies in anticipation of the home-grown resources that are likely to emerge. By demonstrating that "traps of the cure" can be avoided, it exemplifies how universities can build on a phenomenon occurring in many eastern and central European countries: a burgeoning popular interest in higher education. ■





World pandemic
NIH's Tony Fauci
pledges international
approach to AIDS
p628



Stem cells
France moves to end
ban on human
embryo research
p629



In accord
US medical schools
seek harmony over
conflicts of interest
p630



Road rage
Japanese court rules
that truck emissions
cause illness
p631

Successes in fight to save ozone layer could close holes by 2050

Mark Schroppe

The international effort to halt ozone depletion through legislated reductions in the use of chlorofluorocarbons (CFCs) has been a success, according to leading climate researchers who met near Buenos Aires last month.

The researchers, part of the World Climate Research Programme's project on stratospheric processes, predicted that the polar ozone holes should close for good in about 50 years. But they warned that the complicating effect of greenhouse gases could lengthen the process.

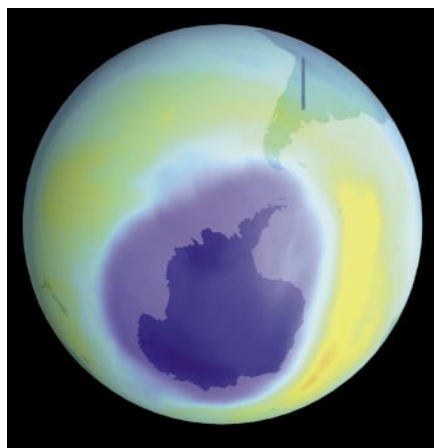
The researchers say there is no direct evidence that the ozone layer is recovering. But, they add, the stabilizing of CFC concentrations in the lower atmosphere means that such a recovery is assured.

Sherwood Rowland of the University of California at Irvine, who shared the 1995 Nobel Prize for Chemistry for his part in the discovery that CFCs destroy ozone, agrees with the researchers' projections, and adds that the effects of greenhouse gases should be relatively small and short lived. "The story is approaching closure, and that's very satisfying," Rowland says.

In the upper atmosphere, or stratosphere, CFCs can spark a complex chain of chemical reactions that leads to the production of chlorine and the destruction of ozone. Researchers say that CFC concentrations will take another decade or so to level off and begin to decrease — indeed, the biggest ever ozone hole was seen over the Antarctic earlier this year. The reason for the time lag is that CFCs are inert and persist in the atmosphere for up to a century.

This delay has led to some confusion about the effectiveness of the CFC bans agreed by 170 countries in the 1987 Montreal Protocol. But Alan O'Neill, a climate modeller at the University of Reading in England and chair of the meeting's organizing committee, says that the record-breaking hole is not surprising.

O'Neill says that models of how the atmosphere should behave now match observed trends well. Assuming that current CFC policies continue, the models forecast



Hole story? The ozone hole looms large in this September image from the TOMS-EP satellite.

that the ozone holes that open up over Antarctica each year and over the Arctic periodically should be gone by about 2050.

But he says that higher concentrations of greenhouse gases — the effects of which have not been reliably modelled — could push that date back by a few decades.

Greenhouse gases are thought to impede ozone repair by expanding the area of the stratosphere in which ozone is destroyed. This is because the gases, although they warm the lower atmosphere, cool the stratosphere, forming ice crystals which catalyse the destruction of ozone by CFCs.

The combination of lingering CFCs and more greenhouse gases could lead to even more severe ozone holes in the short term. "Ozone could be at its lowest ever in five years time," says O'Neill, "but in the longer term we ought to be in business for recovery." Large volcanic eruptions, which launch dust into the stratosphere, could also cause short-term depletion, researchers say.

Drew Shindell, an atmospheric chemist at the NASA Goddard Institute for Space Studies in New York, says that the meeting marked a new appreciation of the confounding role of greenhouse gases in ozone depletion. Nonetheless, he says, "the agreement to limit production [of CFCs] looks like it's been an unqualified success. The science was listened to, policy-makers did something, and it actually worked."

Gridlock stalls NIH budget rise

Paul Smaglik, Washington

While the United States enters December without having elected a president, biomedical researchers face an even more pressing problem. Thanks to gridlock in Washington, the National Institutes of Health (NIH) still has no budget for the 2001 financial year, which began on 1 October.

Congress returns this week for a session that was expected to agree on the appropriations bill that funds the NIH. The bill was expected to include a massive 15% increase, taking the agency's budget to \$20 billion. But with Congress disinclined to pass a version that outgoing president Bill Clinton would sign, it is anyone's guess when the NIH will get the money.

As the first grants for the 2001 fiscal year go out this week, the uncertainty is starting

to hit home, says Wendy Baldwin, the NIH's deputy director of extramural research. "Not as many people will get money, and those that will won't get as much."

Even if the NIH gets its 15% increase, there is no guarantee that the money will be backdated to October. If the electoral impasse leaves the NIH with a smaller increase than had been expected, it would dampen the agency's chances of doubling its budget over a five-year period — the full 15% would take the NIH halfway there.

Biomedical research lobbyists are unsure how or when the impasse will end. At worst, the government might maintain the 2000 levels into 2001, and then decide that the agency could not manage a year's worth of budget hikes in a shortened period.

► <http://grants.nih.gov/grants/news.htm>

US puts global collaboration at heart of AIDS effort

Rex Dalton, San Diego

The US National Institutes of Health (NIH) has announced a global strategy for fighting AIDS that will place more emphasis on international research and scientific training.

Coordinated by the NIH's Office of AIDS Research (OAR), officials say that the plan is expected to be the basis for greater US involvement in international collaborations.

The NIH spent about \$90 million last year on international AIDS research, with that amount expected to increase to \$100 million when Congress approves the delayed 2001 federal budget (see page 627).

Officials cannot yet discuss their funding plans for the initiative in 2002, but say that the growing concern in the United States about the international impact of AIDS should lead to "significant increases".

"A strong, coordinated biomedical research effort is critical to address the global pandemic," says Jack Whitescarver, OAR's acting director.



NIH's Anthony Fauci backs international AIDS research.

The strategic plan is expected to encourage training for international scientists to take part in NIH studies, research workshops, scholarships for international scientists, prevention strategies and new approaches to research funding.

The NIH developed the plan with a panel of outside academics, and

released it on 1 December to mark World AIDS Day. "It is our sincere hope that the fruits of this research will help to alleviate the suffering caused by HIV and AIDS throughout the world," said Anthony Fauci, director of the National Institute of Allergy and Infectious Diseases.

Advocacy groups for AIDS research reacted cautiously to the announcement, noting that it is largely a repackaging of existing programmes. One said that he did not want to comment publicly on the plan, as the NIH was trying to heighten awareness of international issues and it was too early to tell whether the initiative would lead to substantial change.

Officials at another advocacy organization, the New York-based Treatment Action Group (TAG), said that the plan may plug gaps in research. For instance, says TAG's Mark Harrington, the United States now spends about \$50 million a year on research related to US children with AIDS, but there are only a few such cases. Globally, he points out, paediatric AIDS is a huge problem, but the United States probably spends less than \$5 million each year researching it.

In the early 1980s, the US government began international AIDS research projects in Haiti and Zaire. It then expanded AIDS research projects to 50 countries in Africa, Asia, Europe and Latin America.

Officials say that the new plan will help researchers to build new collaborations. The NIH will form a Global Strategy Group to track the plan's implementation.

◆ nih.gov/od/oar

Scientists win vote of confidence from French public

Declan Butler, Paris

People in France still see science in a positive light despite the recent tempests over mad cow disease, HIV in blood, and genetically engineered food, according to a survey undertaken for the research ministry.

Of 1,000 adults surveyed by the polling organization Taylor Nelson Sofres, 53% said they most trusted scientists to control the progress and ethical use of science — far ahead of intellectuals and philosophers (19%), associations and trade unions (9%), or religious groups (6%). Politicians were most trusted by only 4%.

Roger-Gérard Schwartzberg, the minister for research, said the results showed an urgent need to incorporate science more effectively into decision-making. There should be more debate of contemporary scientific issues, he said, and political parties and presidential candidates should increase the discussion of science in their manifestos.

When respondents were asked to choose priority areas, medical research came out on top, with 84% of the vote, followed by environmental science at 54% and energy research at 32%.

Nine-tenths of those polled said research should be a priority for France, and two-thirds that the research budget should be increased; only 3% thought it should be reduced. A career in science was regarded as attractive by 84%.

Mars images may be legacy of lakeside view

Tony Reichhardt, Washington

NASA seems to have trouble keeping secrets, when it comes to Mars.

Four years ago, it was claims of fossil life in a Martian rock that set the media in a spin, after word leaked out ahead of publication in the journal *Science*.

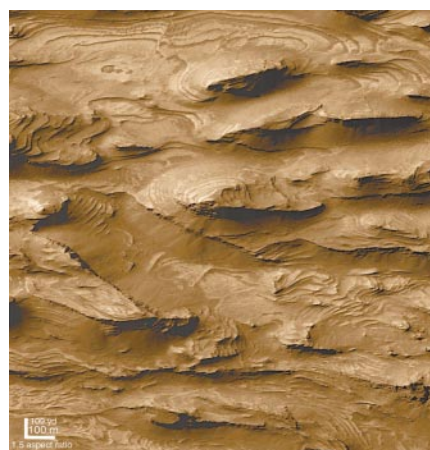
Now, in an uncanny echo of these events, a NASA press conference originally planned for Thursday 7 December was hastily pulled forward to the Monday after news services ran the story that scientists Michael Malin and Ken Edgett of Malin Space Science Systems in San Diego had found evidence for ancient lakes on Mars.

Malin and Edgett report in the 8 December issue of *Science* on layered outcrops that appear to be sedimentary rocks in several natural basins. The outcrops, which are

between 4.3 billion and 3.5 billion years old, are possible evidence that surface water was widespread in the planet's distant past.

The researchers caution that water may not be the only explanation for the images returned by NASA's Mars Global Surveyor. It is also possible, for example, that atmospheric pressure was once higher on Mars than it is today, causing winds to spread dust around the surface.

Layered deposits were discovered on Mars in the 1970s, but have never been seen in so much detail. The lake sediments — if the evidence holds up — would have important implications for any future fossil hunts. "These discoveries give us some direction on where to go," said Jim Garvin, NASA's Mars exploration.



Splash down? These layered patterns on the surface of Mars could have been caused by water.

NASA/JPL/MALIN SPACE SCIENCE SYSTEMS

NIH cash tied to compulsory training in good behaviour

Rex Dalton, San Diego

A sweeping new government policy requiring US researchers and other laboratory staff to undergo training in the "responsible conduct of research" was issued last week. It will begin to take effect late next year.

Under the policy, institutions receiving grants from the National Institutes of Health (NIH), which supports most biomedical research in the United States, must have a plan in place next October to deliver the training in their laboratories.

Courses that researchers will have to take include instruction on ethical scientific conduct, work with human and animal subjects, conflicts of interest, and authorship rules. The health department's Office of Research Integrity (ORI), which drew up the policy, says that a three-hour course, available on compact disc, can meet its requirements. But more comprehensive courses, including group dialogues on case studies, are expected to become the norm.

By October 2003, all personnel involved in research — principal investigators, post-docs, students and lab technicians — must have been trained to conduct studies under a Public Health Service (PHS) grant (the NIH is part of the PHS). Institutions or laboratories that fail to implement the training face losing all PHS research funds.

Chris Pascal, the ORI's director, says he hopes his agency has struck an appropriate balance in the new training policy. "We don't believe institutions will fight it," he says.

But when researchers learned of the development last week, they expressed some concern about the time the training would take. At the Salk Institute in La Jolla, California, virologist Matthew Weitzman said that junior principal investigators such as himself may consider the training unnecessary. "As junior investigators, we are already overburdened," he says, adding that the training may be perceived as just another burden.

But neuropharmacologist Michael Kalichman, director of the research ethics programme at the University of California at San Diego, says he expects the new training courses to win converts at all levels — as have voluntary courses in research ethics.

For Sheldon Krinsky, a professor of urban and environmental policy at Tufts University in Massachusetts, the new policy represents "an important first step" in acknowledging the problem of scientific misconduct. Proper ethics "have to be built into the scientific culture," says Krinsky.

The ORI, which monitors scientific conduct for NIH grants, issued the policy



after several months of discussions with universities and institutions that had followed several years of informal debate.

The National Science Foundation, which supports most non-biomedical research at US universities, is not planning to introduce similar training. But, historically, other research agencies in the US government have sometimes followed the ORI's lead on issues related to scientific misconduct.

For nearly a decade, the ORI and associated government agencies have relied on research institutions to monitor, investigate and report on the conduct of scientists receiving grants. The ORI oversees this process, which can lead to penalties — from warnings to debarments — for misconduct.

High-profile incidents in recent years — including instances of mistakes in clinical trials, researcher conflicts of interest and data fabrication — have prompted the government to add the more stringent training requirements in a bid to strengthen scientific ethics, officials say.

Although some universities are already planning their own training programmes on the conduct of research, only a handful of the 300 major research institutions operate them now. Some of those that do, such as the University of Minnesota, set them up as part of a remedial plan after particular instances of research misconduct.

The new policy was welcomed by leaders of research institutions and by organizations that had criticized earlier ORI proposals. An official at the Council on Governmental Relations, which represents leading research universities, called the proposal "well-researched and thoughtful". Some university representatives had rejected the training proposal as overly prescriptive and bureaucratic. But concessions by ORI on the timing and extent of the training proposal appears likely to blunt university opposition. ■

► <http://www.ori.hhs.gov>

France opens door to use of embryos in stem-cell research

Declan Butler, Paris

The French government is to submit a bioethics bill to parliament that would lift its ban on human embryo research.

The decision was widely expected, but some observers are surprised that the bill does not expressly prohibit therapeutic cloning of human embryos to create embryonic stem cells. Last month, advisers to the European Union concluded that such a move would be "premature" (see *Nature* 408, 277; 2000).

But the bill will state that it should "not be excluded *a priori*", says Roger-Gérard Schwartzberg, the research minister and chief sponsor of the measure, as it may become necessary should other techniques fail. He points out that cells or tissues derived from embryonic stem cells produced by cloning for transplantation would not be rejected by the patient's body.

The bill — an update of France's 1994 bioethics legislation — would allow research on embryos left over from *in vitro* fertilization procedures. Research would be restricted to embryos less than 6–7 days old and to circumstances in which no effective alternatives existed.

Protocols would be evaluated and overseen by a proposed new government-appointed body. This would be responsible for human reproduction, developmental biology and genetics, and predictive medicine.

The bill has been drafted largely by the pro-science research ministry. Government officials predict that it will face a rough ride through the conservative senate, and perhaps even the parliament, where a Socialist-led coalition holds a majority.

Opponents will argue that creating embryos for research runs counter to the European convention on bioethics, which France has signed, and that the proposed therapeutic cloning provisions would prevent France from ever ratifying the convention.



Research minister Schwartzberg will not rule out cloning.

Others say that the bill is calculated to defer difficult decisions. "The new authority will rule on a case-by-case basis, meaning that decisions will be postponed until four or five years down the line," says one official. ■

AFP

Medical schools in concert on research ethics

Steve Nadis, Boston

Back in May, officials at Harvard Medical School gingerly stepped back from a plan to relax the institute's conflict-of-interest policies, which are considered among the strictest in the United States.

At the time, they called for a national forum to establish more uniform standards of conduct for researchers at US academic medical centres (see *Nature* **405**, 497; 2000). Last week, a closed meeting in Washington edged these standards a step closer to reality.

Organized by Harvard's dean of medicine, Joseph Martin, the meeting brought together representatives from eight of the ten largest medical schools in terms of funding from the National Institutes of Health (including Harvard, Yale, Johns Hopkins, and Washington University). Also present were high-profile leaders such as former National Institutes of Health director Harold Varmus, now president of the Memorial Sloan-Kettering Cancer Center in New York.

According to those who attended, the meeting focused on the conflicts of interest that investigators might face when their research offers the potential for personal gain. Recent studies, including one that appeared last week in the *New England Journal of Medicine* (**343**, 1621–1626; 2000), have shown a high degree of variation in the ethics



Raising the standard: Harvard's Joseph Martin wants a common policy for conflicts of interest.

policies of medical schools and teaching hospitals. Some have relatively strict rules, whereas others have no stated policies at all.

"This group agreed that it would be good for our community if there were less variability in these policies, while also encouraging the policies to be strengthened overall," says one attendee, David Korn, vice-president for research at the Association of American Medical Colleges (AAMC). "There was also agreement about the need for absolute and full disclosure, both internally and externally, regarding the possible financial interests of investigators," he adds.

The group agreed that every institution should have explicit conflict-of-interest policies, clear standards for disclosure of

financial ties, and mechanisms for ensuring compliance with the guidelines. "These are broad principles, not detailed rules," Martin says. "We don't get into the nitty-gritty, as those details should be worked out by the individual institutions."

A draft statement of principles is now being circulated among meeting participants and could be released before the end of the year, says Dennis Kasper, dean of academic programmes at Harvard Medical School. "We don't represent anyone other than a small consortium of schools that met to talk about this, but we hope other schools will consider these principles."

Martin says that other institutions may be brought on board through the AAMC. The association has appointed a task force, chaired by William Danforth, former chancellor of Washington University, to review conflict-of-interest issues next year.

And although major medical schools are each likely to adhere to their own sets of conflict-of-interest rules, administrators believe that some commonality between these rules will help to discourage researchers from 'jumping ship' to more lenient institutions.

But Korn warns that there are more important issues at stake. "The critical thing is to maintain the public's confidence in medical research," he says. ■

GRAHAM RAMSAY/AERO PHOTO

Xenotransplantation opponents take FDA to court

Paul Smaglik, Washington

Opponents of xenotransplantation research have begun a lawsuit against the US Food and Drug Administration (FDA), demanding that more information be made public about experiments in which humans receive transplants of animal tissue.

The Campaign for Responsible Transplantation, which filed the lawsuit, says adverse events from xenotransplantation have been reported in journals and at public meetings hosted by the FDA, but the agency has refused to release all the data from these trials. An FDA spokesperson declined to comment on the lawsuit.

The campaign group compares the work's secrecy to that which surrounded gene therapy until the death of a patient, Jesse Gelsinger, during clinical trials at the University of Pennsylvania last year. Alix Fano, director of the group, says that it wants xenotransplantation research to be as open as gene-therapy trials now are. "This is a new form of biotech — it's akin to the gene-therapy trials where a lot of side effects went unreported," she says.

Fano estimates that there are around a

dozen xenotransplantation trials under way in the United States. The precise number is unknown, she says, because the FDA does not disclose it. The trials began more than 30 years ago, with about 12 transplants of chimpanzee organs. More recent experiments have studied the transplantation of cells, rather than complete organs.

LeRoy Walters, director of the Kennedy Institute of Ethics at Georgetown University in Washington DC, supports the idea that data from medical research should be more freely available. "I would like to see a public registry of all clinical trials across the board," he says.

The lawsuit reflects the tension between the industrial sponsors of research, who want to keep the results of clinical trials private, and some public-health advocates, who want more openness. That tension increased last autumn following Gelsinger's death, which set in motion a chain of events that resulted in a publicly accessible database of gene-therapy clinical trials.

Drug companies say that the release of information could threaten their competitiveness. But campaigners and



Contained: data on adverse events in animal transplant studies have not been freely released.

bioethicists argue that both gene therapy and xenotransplantation differ so much from conventional medicine that more transparency is needed.

Gene therapy, they say, runs the risk of germline gene transfer, where genetic changes are passed unintentionally to the patient's children. With xenotransplantation, critics worry that retroviruses might pass from animals to humans. ■

AP

Japan makes polluters pay after landmark court ruling

David Cyranoski, Tokyo

The clean-air movement in Japan scored a significant victory last week when a court judged that vehicle exhaust and industrial emissions had adverse effects on human health. Scientific evidence was pivotal in the court's ruling.

Ending an 11-year court battle, the Nagoya district court ordered the Japanese government to pay ¥18 million (US\$167,000) and a group of ten industrial companies to pay ¥289 million to victims of pollution.

Although small in financial terms, the ruling is likely to have far-reaching implications for Japan's air-pollution policies. The court also directed the government to enforce policies that will keep the concentration of particulates in the air to below the limits set by Japan's Environment Agency.

The Ministry of Health confirmed that the plaintiffs were suffering from pollution-related diseases. And the court, after hearing epidemiological evidence, judged that particulates in the air could lead to such illnesses as "life-threatening bronchial asthma".

Two studies were particularly influential, observers say. One was a ten-year epidemiological study by the School of Public Health at Chiba University on the rates of asthma in schoolchildren living close to major roads. According to Masayuki Shima, a researcher on the study, their results showed that the likelihood of developing asthma was 3.7 times higher than average in boys and 5.9 times higher in girls who lived within 50 metres of busy roads. The researchers at Chiba attribute the difference to particulates from diesel exhaust.

In the second study, the Environment Agency's National Institute for Environmental Studies (NIES) found a clear connection between particulates and animal health. When mice were made to inhale diesel exhaust, they developed asthmatic symptoms and suffered damage to their reproductive organs. The results of five years' research were released last year in a NIES report.

The Environment Agency and the ten companies have questioned the court's interpretation of these results. One Environment Agency official says: "There are so many things that are still uncertain. It is quite possible that the effect on humans is different from that on mice. Likewise, it is hard to say that the high levels of exhaust given to the lab animals are anything like those experienced by people in real life."

Masaru Sagai of Aomori University of Health and Welfare, who led the research team at NIES, responds: "It is necessary and scientific to extrapolate from experimental results—for example, from the effect of high doses over a short time to the effect of low doses over a long time."

Environmentalists say these problems would stand a better chance of resolution if the government were to invest seriously in environmental research. Compared with the United States, in particular, "Japan spends very little on researching and monitoring the effects of pollution", says Takao Nishimura, an environmental lawyer who has been involved in many such cases.

And there is some concern that successful lawsuits may diminish the government's interest in supporting pollution research. Last January, a similar ruling was given in Amagasaki, and in the months that followed, research projects allocated to Chiba University by the regional governments of Tokyo and Chiba were cut off.

Critics allege that the funding was withdrawn because of the governments' fears about the power such data would give people suing government agencies. But Shima says he has no reason to doubt that the cuts were made for the official reason given, namely the financial straits of the sponsors.

Even with financial support, says Sagai, his research has had little effect. "Once it became clear that my research on mice produced results that were disadvantageous to the agency, they ignored it altogether."

http://www.nies.go.jp

Breathe, don't breathe: a Japanese court has linked traffic fumes with childhood asthma.



© 2000 Macmillan Magazines Ltd

Academics bid to transcend the Arab–Israeli conflict



Brave face: but confrontation is taking its toll.

Haim Watzman

Joint research projects between Israeli and Arab scientists are continuing despite the current confrontation, but at a reduced level, scientists in the region say.

A study last summer counted 132 projects in the natural sciences involving collaboration by researchers from Israel and the Palestinian authority or other Arab countries.

There has been no comprehensive survey on the status of these projects since the outbreak of Israeli–Palestinian violence in September. But although some activities have been cancelled and military restrictions on travel have made it difficult for Palestinian students to reach Israeli campuses, researchers say that individual cooperation continues as much as possible.

Several international scientific conferences scheduled to take place in Israel have been delayed or switched to other venues. The most important was that of the Red Sea research programme, an oceanographic and environmental endeavour involving scientists from Israel, Palestine, Jordan and Egypt.

Immediately after the outbreak of violence, some scientists lost touch. Moein Kan'an, for example, a researcher at Bethlehem University who is involved in several projects with Israeli scientists, says that for the first two weeks of the conflict he heard nothing from his Israeli research partners. "We had a serious talk and resolved that we were going to proceed," says Doron Lancet of the Weizmann Institute of Science, who is working on a project with Kan'an.

The Israel Academy of Sciences and Humanities held a day-long seminar last summer to evaluate scholarly cooperation between Israelis and Arabs. Many Israeli scholars voiced their belief that science was one field in which this was increasing, although others expressed doubts.

Europe's patent office skips debate on biotech rules

Munich Member states of the European Patent Office (EPO), meeting last week, disappointed both researchers and opponents of gene patenting when they declined to clarify European rules for the protection of discoveries in biotechnology.

"Considering the European Union's leading political and legislative role in this area, the EPO decided not to start a parallel discussion," said Roland Grossenbacher, president of the EPO.

The meeting did agree that computer programs should not be patented. Critics say this rule drives software companies from Europe to the United States or Japan, where programs can be patented.

The EPO also agreed on steps to increase the efficiency of the European patent system, such as allowing patent applications to be filed in any language.

Japan targets environment and nanotechnology

Tokyo Japan's Council for Science and Technology, chaired by Prime Minister Yoshiro Mori, has released a draft five-year plan for

science and technology, giving top priority to research in the life sciences, the environment, nanotechnology and information technology.

Research funding for 2001–2005, the period covered in the plan, will total ¥24 trillion (US\$216 billion), including a doubling of competitive grants from ¥300 billion to ¥600 billion each year.

The plan will also push for mobility among young researchers in government laboratories by putting those under 35 on fixed-term five-year contracts. It will also allocate money for a more thorough system of research evaluation, involving more external and international review of projects.

Coelacanth stars in South African film

Cape Town The coelacanth, an ancient fish that has existed for at least 360 million years, has been filmed for the first time in decades by divers exploring a reef at Sordwana Bay, just south of South Africa's border with Mozambique.

Michael Bruton, secretary of the Coelacanth Conservation Council, says there is absolutely no doubt that the pictures are of a coelacanth. The species, which was thought to be extinct until the last sighting in South African waters in 1938, has also been observed off Madagascar and Mozambique.



Caught on camera: the ancient South African coelacanth was considered at risk of extinction.

In 1997, a second species of coelacanth was discovered near the Indonesian island of Manado Tua.

Germany looks to sequence bacterial genomes

Munich The German government has launched an initiative in microbiology that will concentrate on investigating bacterial genomes.

Under the GenoMik initiative, the federal science ministry plans to spend DM40 million (US\$18 million) over the next three years studying these genomes. It hopes the work will lead not only to treatments for human disease, but also to a better understanding of bulk processes that may

have potential applications in industry or for cleaning up the environment.

The initiative completes a portfolio of federal funding for research on human and plant genomics. Both academic and industrial institutions are eligible to participate.

Royal Society calls for better communication

London Britain's Royal Society is to spend £1 million (US\$1.5 million) on encouraging scientists to communicate more effectively with the public.

Details of the Science and Society programme remain to be thrashed out, but it will be headed by Paul Nurse, director-general of the Imperial Cancer Research Fund. He has asked outside consultants to evaluate and cost different ways of communicating science to the public. Their report is expected later this month.

Projects may include communication training for scientists, awards for researchers who explain their work to a wider audience, and links between science museums and research agencies. The five-year initiative was announced last week by the society's outgoing president, Aaron Klug, in his farewell address, and will be paid for by a grant from the Kohn Foundation.

Biomedical whistleblowers to receive protection

San Diego New rules to protect 'whistle-blowers' who inform on people suspected of research misconduct have been proposed by the US government.

The rules will require institutions receiving grants from the Public Health Service — which includes the National Institutes of Health — to produce written procedures aimed at preventing retaliation against whistleblowers who expose scientific or financial misconduct.

Interested parties have 60 days to comment on the rules, which were drawn up by the Office of Research Integrity.

► <http://www.ori.hhs.gov>

Large celestial object stuns astronomers

Washington The Solar System gained a new associate member last week, when astronomers in Tucson, Arizona, found an object that may be half as large as Pluto in the Kuiper belt, beyond the orbit of Neptune.

The object, now designated as WR106, was spotted by the Arizona Spacewatch team during routine scans for potentially dangerous objects crossing the Earth's orbit. Brian Marsden of the Harvard-Smithsonian

Center for Astrophysics in Massachusetts estimates that it is about 1,000 kilometres in diameter.

WR106 is so bright that Marsden believes it will show up in photos taken over the past 40 years by the Palomar Sky Survey. These data, together with further observations over the next two months, should allow astronomers to determine its orbit.

HIV-positive men offered hope of healthy children

Tokyo In an experiment that could allow men with AIDS to have healthy children, doctors at Niigata University plan to perform *in vitro* fertilization this month using HIV-positive sperm from which the virus has been removed. The virus-removal method, developed by researchers at Ogikubo Hospital and Keio University, involves breaking down viral RNA or DNA within the semen and using a centrifuge to recover the sperm.

Researchers say the method will allow the selection of sperm with a less than one in a 100,000 chance of being tainted by HIV. They add that, even if the embryo happened to be infected, the use of *in vitro* fertilization decreases the chances of the mother being infected. The experiment has been approved by the ethics committees of the relevant universities.



The great ice mystery

Changes in the extent and thickness of sea ice could alter ocean circulation and so disrupt the climate. Jon Copley considers one of the big unknowns in the global warming debate.

Contrary to popular myth, the Inuit do not have dozens of words for snow. Oceanographers, however, do have a multitude of terms for sea ice — including frazil, grease, shuga, pancakes and brash. Each describes a different stage in the freezing of surface sea water.

This shifting frozen skin is a crucial boundary between the oceans and the atmosphere. It regulates the transfer of heat, and gases, and is a vital part of the 'engine' that drives the circulation of the world's oceans. Given that changes in this circulation could have profound effects on climate — plunging Western Europe into the deep freeze while most of the world warms, for instance — a better understanding of the dynamics of sea ice is becoming a top priority.

Fortunately — thanks in part to the declassification of data gathered by military submarines during the Cold War — the necessary information is starting to come to light. Even so, researchers investigating sea ice have their work cut out, if they are to assist the modellers who are trying to anticipate the effects of climate change — and so help politicians decide how to respond to the challenges ahead. "We now have a fairly massive data-processing task," says Peter Wadhams of the United Kingdom's Scott Polar Research Institute in Cambridge.

Salt water has a lower freezing point than fresh, so sea ice usually begins to form when the temperature of surface water drops below -1.8°C . This process is quite distinct from the creation of icebergs, which happens when glaciers meet the sea. At first, tiny crystals known as frazil appear; these then coagulate into a soupy slick known as grease. If conditions are calm, grease ice can grow into a thin, elastic crust called nilas, which may thicken into more substantial cover. More often, wind and waves push grease ice into pancake-shaped chunks that eventually bond to form floes.



Chill out: the bestiary of ice ranges from the small ice flowers that form on sea ice (top) to the large chunks of pancake ice that bond to form floes. The dynamics of sea ice are now coming to light.

Sea ice may continue to congeal on the underside of existing floes, or even form on their upper surfaces if the the weight of accumulating snow pushes the floes below sea level. Colliding floes can be squeezed into thick ridges. And although much of the ice melts away each summer, some survives to accumulate new growth in subsequent years. As a result of all these processes, sea ice can take a multitude of forms — hence the exotic names devised by its aficionados.

Poles apart

The yearly cycle of freezing and thawing differs between the poles. The largely landlocked Arctic Ocean offers fewer opportunities for sea ice to drift to warmer latitudes, and so retains a greater proportion of multi-year ice. Melting in the Arctic tends to begin at the surface, producing pools of melt water that may absorb more

heat and promote further melting. But melt pools are rare in the Antarctic, where melting ice can more easily break up, exposing areas of open water. This allows greater heating of the surface ocean, which melts the sides and bottom of remaining floes. Little ice survives more than one year in the waters around Antarctica, and as a result the seasonal extent of the ice is much more variable than in the north. Consecutive satellite images of sea ice around the continent reveal a pulsating halo that waxes and wanes with the seasons.

In addition to this seasonal cycle of growth and decay, the extent of sea ice seems to vary over longer timescales. In recent years, there have been several signs that sea ice has receded in the Arctic. Data from satellite passive-microwave observations — which, by recording thermal radiation from the Earth's surface, allow sea ice to be distin-

guished from open water — reveal that between 1978 and 1996 the average overall extent of Arctic sea ice shrank by around 34,000 square kilometres per year, or 2.8% per decade¹. The season during which there is a net melting of sea ice also appears to have lengthened by 8% per decade over the same period².

The extent of ice is not the whole story, however. “Satellites see ice extent very well, but it’s a little harder for them to figure out how thick the ice is,” says Drew Rothrock of the University of Washington in Seattle. But the submerged depth of ice, or ice draft, can be measured from below using sonar. To get these data, Rothrock has been on Arctic cruises aboard US Navy submarines as part of a programme called Scientific Ice Expeditions (SCICEX). The programme also has access to recently declassified ice-draft data collected by submarines during the Cold War.

SCICEX cruises in the 1990s measured a mean ice draft of 1.8 metres above the deep waters of the Arctic north of Canada at the end of each melt season. In contrast, submarine patrol data from 1958 to 1976 show that the mean used to be 3.1 m (ref. 3). “It’s a terrific record and it’s only just being made public,” says Rothrock. “We’re trying to get more data into the public archive.”

Thinning on top

Meanwhile, British submarines have charted a complementary record of changes in ice thickness in other regions of the Arctic. “We’re finding the same amount of thinning in summer on the European side as Rothrock found on the Canadian side,” says Wadhams, who cruised in Royal Navy submarines in 1971, 1976, 1987 and 1996⁴. This year, the UK Hydrographic Office released data on cruises from 1988 to 1995, filling gaps in Wadhams’s record. He is now scrutinizing the record to see whether there was steady thinning over the period, or whether most of it happened in one go.

The big question is whether the changes in Arctic sea ice are a consequence of global warming. General circulation models (GCMs), which modellers use to probe the effects of rising levels of greenhouse gases, predict that global warming should be enhanced in the Arctic. This may be amplified by a mechanism known as ice-albedo feedback. Sea ice is highly reflective, particularly when it is covered with snow. So less sea ice means that less solar radiation is reflected, and the sea absorbs more energy, enhancing local warming and promoting further melting. Disappearing sea ice may therefore hasten its own demise by exposing more open water to the Sun.

This positive feedback can work the other way, with growing sea ice promoting the formation of yet more ice — particularly when it extends far from the poles. Indeed, during

the late Proterozoic era, between 600 million and 800 million years ago, such a runaway effect may have encased most of the planet in ice — creating a ‘snowball Earth’⁵.

Sea ice cover may oscillate naturally over timescales of decades or longer. But researchers do not know whether the present decline in the Arctic can be explained by such oscillations or whether it is a direct result of the build-up of greenhouse gases in the atmosphere. Another possibility is that the changes in ice cover are being driven not by a slow increase in air temperature, but by changes in ocean circulation. “It’s possible that most of the thinning happened from 1990 onwards, because that’s the time when there’s been a change in the oceanographic conditions in the Arctic,” says Wadhams. In the past decade, he points out, there has been more warm water flowing into the Arctic from the Atlantic.

Meanwhile, at the other end of the world,

the jury is still out on what effect increased atmospheric carbon dioxide will have on sea ice. Some GCM simulations predict a very slow warming of surface air⁶, but others predict more rapid changes that would cause the extent and thickness of sea ice to be drastically reduced⁷. So far, satellite data show no signs of a reduction in the overall extent of Antarctic sea ice.

View from the bridge

But down south, it is hard to monitor ice thickness, because the Antarctic Treaty forbids military submarine cruises. To fill in the picture, the Antarctic Sea Ice Processes and Climate (ASPeCt) programme of the Scientific Committee on Antarctic Research is compiling observations of sea ice made from ships between 1980 and 1997⁸. Because different categories of sea ice can be used as a proxy for ice thickness, observations of ice type — typically made every



Packed in: the *Des Groseilliers* spent a year frozen into the Arctic ice as a research station.

hour from a ship's bridge — can be translated into estimates of ice thickness. So far, the ASPeCT data comprise 11,000 observations from 42 voyages.

The programme expects to release the first circumpolar distribution of sea ice thickness compiled from these records early next year. As part of the ASPeCT programme, Tony Worby of the NASA Goddard Space Flight Center in Greenbelt, Maryland, has produced a CD-ROM to train observers to make consistent records during future Antarctic voyages. These records will be added to an archive at the Antarctic Cooperative Research Centre in Hobart, Tasmania, to provide a long-term time series of ice thickness estimates.

Programmes such as ASPeCT and SCICEX should, over the next few years, allow researchers to begin to understand how global warming is affecting sea ice cover. But if sea ice is receding, what are the consequences?

Failing circulation

For commercial shipping, a reduction in ice cover could bring benefits. This summer, the Royal Canadian Mounted Police patrol vessel *Nadon* (temporarily rechristened the *St Roch II*) made a transit through the Northwest Passage — the shortcut between the Pacific and the Atlantic across the top of North America — and encountered very little sea ice along the way.

But changes in ice cover could have far-reaching effects on climate. When surface sea water freezes, its dissolved salt is left behind, forming a cold, dense brine that tends to sink. This mixes with the water below in a process known as deep water formation, which is an important part of the thermohaline circulation — the conveyor belt of ocean currents that spreads heat around the globe and has a strong influence on climatic conditions. Indeed, some researchers believe that sea ice variability may cause the switch between ice ages and interglacial periods.

Wadhams is involved in a project over the next three years to examine the interaction between sea ice and the thermohaline circulation in the Greenland Sea. In this area, the Odden ice tongue — a seasonal sea ice feature that sticks out from the coast of Greenland — has failed to form in five of the past seven years⁹. If the loss of the ice tongue is slowing the formation of deep water, it could seriously disrupt the thermohaline circulation in the North Atlantic. Potentially, it could weaken the flow of warm water from the Gulf of Mexico that currently ensures that northwest Europe enjoys a much milder climate than its latitude would otherwise allow.

Modellers have only recently begun a concerted effort to examine the effects of changing sea ice cover on the thermohaline



On thin ice: one of the *Des Groseilliers*'s crew measures the extent of melting in the Arctic.

circulation. One of the first of these studies suggests that the extent of sea ice and the strength of the thermohaline circulation could oscillate gently over the decades¹⁰. Other studies indicate that there may be destabilizing feedbacks that could weaken the thermohaline circulation^{11,12}. Jochem Marotzke of the Southampton Oceanography Centre, a co-author of one of these studies, believes the models must be improved to make better use of existing knowledge of sea ice processes.

Go with the floe

To check the validity of existing models, the National Science Foundation and the Office of Naval Research funded an experiment called SHEBA — the Surface Heat Budget of the Arctic. Led by Don Perovich of the US Army's Cold Regions Research and Engineering Laboratory in Hanover, New Hampshire, the SHEBA team let the Canadian icebreaker *Des Groseilliers* freeze into the Arctic ice pack to provide the headquarters for an ice station.

The ship and the cluster of tents around it drifted 640 kilometres across the Arctic with

the pack ice for a year from October 1997. During this period, the ice station's crew conducted an intense series of atmospheric, oceanographic and ice observations over a floe that initially measured roughly eight kilometres square — chosen to be about the same size as the individual computational cells that make up a typical GCM.

One of the SCICEX submarine cruises also buzzed the SHEBA site during the experiment to check the team's measurements of ice thickness. One of the key findings of SHEBA was that the sea ice was less than two metres thick to start with, but ended up even thinner at the end of the experiment, despite undergoing a full annual cycle of growth and melt. Full results from the programme are expected to appear next year in a special issue of the *Journal of Geophysical Research — Oceans*.

The SHEBA team also carried out many studies of small-scale changes in sea ice, which together suggest that the devil may be in the details. Hajo Eicken of the University of Alaska in Fairbanks examined the growth of meltwater ponds on the surface of the ice during summer. Meltwater pools are less reflective than snow-covered ice, allowing them to absorb more heat and grow. But if this causes the formation of channels that allow water to drain through into the ocean, bare ice will be exposed once again. So tiny flaws and channels in the structure of sea ice may be critical in determining the amount of solar energy absorbed by the Arctic in summer.

"This illustrates how the refinement of models involving sea ice depends on results from studies of sea ice and sea ice processes at very small scales," says Eicken. "Our improved understanding of how the plumbing system works may allow us to assess whether this is a process that large-scale climate modellers need to worry about."

As the world's government's grapple with the issue of climate change, those studies could prove crucial. For until the ocean's icy skin reveals the full extent of its subtleties, those attempting to forecast our future climate could be shooting in the dark.

Jon Copley is a freelance writer, and a teaching fellow at the University of Southampton.

1. Parkinson, C. L., Cavalieri, D. J., Gloersen, P., Zwally, H. J. & Comiso, J. C. *J. Geophys. Res.* **104**, 20837–20856 (1999).
2. Smith, D. M. *Geophys. Res. Lett.* **25**, 655–658 (1998).
3. Rothrock, D. A., Yu, Y. & Maykut, G. A. *Geophys. Res. Lett.* **26**, 3469–3472 (1999).
4. Wadhams, P. & Davis, N. R. *Geophys. Res. Lett.* (in the press).
5. Hoffman, P. F., Kaufman, A. J., Halverson, G. P. & Schrag, D. P. *Science* **281**, 1342–1345 (1998).
6. Stouffer, R. J., Manabe, S. & Bryan, K. *Nature* **342**, 660–662 (1989).
7. Rind, D., Healy, R., Parkinson, C. L. & Martinson, D. G. *J. Clim.* **8**, 449–463 (1995).
8. Worby, A. P. & Ackley, S. F. *Eos* **81**, 181–185 (2000).
9. Toudal, L. *Deep-Sea Res.* **46**, 1237–1254 (1999).
10. Yang, J. & Neelin, J. D. *Geophys. Res. Lett.* **20**, 217–220 (1993).
11. Lohmann, G. & Gerdes, R. *J. Clim.* **11**, 2789–2903 (1998).
12. Jayne, S. R. & Marotzke, J. *J. Clim.* **12**, 642–651 (1999).

Island-hopping invaders hitch a ride with tourists in South Georgia

Sir—Southern Ocean islands have long been considered to represent globally significant examples of pristine environments, important for the conservation of intact ecosystems. This is about to change because of environmental pressure arising from an escalation of tourism to the region (15,000 tourists in 1999)¹, marked climate change, and consequent increases in the impacts of invasive species².

Over the past three years, tourism to South Georgia has tripled¹, and most southern islands now receive such visits. This trend is exemplified by the recent publication of an environmental impact assessment of tourism to Marion Island³.

Commissioned by the agency responsible for the island's administration, the South African Department of Environmental Affairs & Tourism, the impact assessment is a response to increasing demands for tourism. Although citing expert opinion that tourism poses considerable threats to the biota, because it increases the likelihood of invasive species arriving and becoming established, the assessment concludes that, in low numbers, tourism may proceed.

Among several important issues raised by the impact assessment is the compatibility of conservation and access. Just as it is becoming clear that multiple use of protected areas commonly has deleterious effects on biota, the demand for it is becoming more vociferous, and in many instances is entrenched in management policy or laws governing the areas concerned.

Thus, Marion Island was declared a special nature reserve under South Africa's Environment Conservation Act, which — in addition to its conservation clauses — also specifies that any citizen of the republic wishing to view such an area may do so, at once creating a potential conservation conflict.

The environmental impact assessment also highlights the substantial mitigation costs, usually borne by the responsible authorities rather than the tourist operators, likely to be incurred should an alien species establish itself as a result of tourist activities.

Such activities fall into a high-risk category, because they tend to involve short visits by numerous people who have frequently visited similar islands. This is likely to mean greater propagule pressure than a single landing of a few people who spend an extended time at one island, and

is likely to facilitate 'island-hopping' by invasive species.

What makes the Southern Ocean islands particularly significant is that, although their isolation means that species have nowhere to go in the face of human pressure, it also means that conservation problems should be straightforward to manage.

Sadly, the latter seems not to be the case, providing a glimpse into the future as the former becomes true of increasingly fragmented mainland habitats.

Perhaps if tourism were limited now, the islands might provide sight of an alternative where conservation enjoys priority over access.

Steven L. Chown*, **Kevin J. Gaston†**

*Department of Zoology and Entomology, University of Pretoria, Pretoria 0002, South Africa

†Biodiversity and Macroecology Group, Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

1. International Association of Antarctica Tour Operators. http://www.iaato.org/tour_stats.html (2000).
2. Bergstrom, D. & Chown, S. L. *Trends Ecol. Evol.* **14**, 472–477 (1999).
3. Heydenrych, R. & Jackson, S. *Environmental Impact Assessment of Tourism to Marion Island* (Department of Environmental Affairs & Tourism, Pretoria, 2000).

Status of Japanese monkeys under debate

Sir—The article "Row over fate of endangered monkeys" (*Nature* **408**, 280; 2000) indicated that Japanese macaques (*Macaca fuscata*) are listed as an "endangered" animal species.

This information is now out of date. The latest World Conservation Union list, released in October this year, indicates that there are not enough data to qualify the species as endangered. The great majority of Japanese monkeys are not currently listed in any of the World Conservation Union's Red Lists.

In fact, the Japanese monkeys regularly come into conflict with humans — for example, they are agricultural pests to many farmers. As a result, the Environment Agency has granted local governments permission to 'remove' some monkeys. The number of monkeys removed each year is estimated to be 5,000 or more.

Neuroscientists in Japan have obtained only a small fraction of these 'removed' monkeys (about 150 per year); the monkeys would otherwise have been killed. We have never asked for monkeys to be removed solely for research purposes.

The use of animals in research has vastly increased our knowledge of human brain function and malfunction. We emphasize that various lines of

experimentation using monkeys, including Japanese monkeys, have played and continue to play a critical role in developing our understanding of the brain.

Kunihiko Obata

The Japan Neuroscience Society, National Institute for Physiological Sciences, Myodaiji, Okazaki 444-8585, Japan

Call for a fairer deal on grant applications

Sir—Most UK researchers on short-term contracts, like myself, are called 'research associates', and cannot be the principal applicant (which is understandable) or even a co-applicant (which is much less understandable) on a grant application.

The Biotechnology and Biological Sciences Research Council has a 'recognised researcher' option, but the other UK research councils have no relevant mechanism at all. When it comes to applications for studentships, a research associate cannot even be mentioned.

This situation is doubly upsetting. First, when research associates come up with a good idea for a grant application, they cannot get any benefit from it. More and more universities, when they advertise for lectureships, ask for applicants with proven ability to attract external funding. Being named as a co-applicant for research grants would be useful in this regard as well as providing recognition for the work we do.

Second, university departments are losing the benefit of potential funding, as research associates are increasingly disgusted by the situation and are becoming reluctant to share their ideas with academic staff. Instead, they wait in the hope of obtaining a position that will allow them to apply for grants on their own behalf — not necessarily in the United Kingdom. Of course, some of us still participate in grant applications, but too often our names are not mentioned and we get no credit when a proposal is funded.

By writing this letter, I hope to reach as many research associates as possible, and urge them to put pressure on their supervisors to raise this matter with the UK research councils. The councils need to change their policies on grant-application eligibility in the interests of all. Young researchers should be acknowledged for their efforts, and deserve encouragement rather than being made to feel that the word 'associate' in their title is meaningless.

Robert Dorazi

Department of Biological Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK

The quantum centennial

One hundred years ago, a simple concept changed our world view forever.

Anton Zeilinger

When Max Planck announced his quantum assumption in his talk at the German Physical Society in Berlin on 14 December 1900, nobody, including himself, realized that he was opening the door to a completely new theoretical description of nature. Quantum physics has had unsurpassed success in explaining many phenomena — from the structure of elementary particles, through the essence of chemical bonds or the nature of many solid-state phenomena, all the way to the physics of the early Universe. To date, all experiments magnificently confirm all quantum predictions with impressive precision.

Quantum mechanics also led to an immense number of technological applications. Modern high-tech developments would have been inconceivable without it — lasers and semiconductors are just two such examples. But most significantly, quantum mechanics changed our view of the world in a way that was completely surprising and had unprecedented depth.

Born in 1858 and educated in Munich and Berlin, Planck became interested in thermodynamics early in his career. In 1894, a very basic problem captured his attention: how to explain the colours emitted by glowing bodies. The classical explanation of the period worked well for the short parts of the light

spectrum, but did not agree with experiments for all wavelengths. Planck had the advantage of close access to the most recent experimental results obtained by Otto Lummer and Ernst Pringsheim and by Ferdinand Kurlbaum and Heinrich Rubens, also working in Berlin, on the spectral distribution of black-body heat radiation emerging from a hole in a box kept at a certain temperature.

Planck eventually found a full explanation, but only after forcing himself “to an act of despair” by assuming that energy can only be exchanged between the light field inside the box and the walls of the container in discrete quanta, multiples of the energy $E = h\nu$, where ν is the light’s frequency and h is now called Planck’s constant. Planck tried unsuccessfully for many years to find an alternative deriva-

tion of this experimentally successful radiation law from other known laws of physics, but he slowly had to accept that he had found something fundamentally new.

The next important step in the early days of quantum mechanics came in 1905, when Albert Einstein introduced his radical hypothesis of quanta of light to explain the photoelectric effect. For a while, this remained the only significant instance of the quantum being taken seriously. Einstein’s hypothesis met with strong objections from his contemporaries, including Planck himself. As late as 1913, Planck, together with his fellow physicists Heinrich Rubens, Walther Nernst and Emil Warburg, wrote in a recommendation letter for Einstein’s election to the Prussian Academy of Sciences: “One should not hold against him too much that in his speculations he might have occasionally overshot the goal, as for example in his hypothesis of the quanta of light.” Ironically, it was this hypothesis that gained Einstein the physics Nobel prize in 1921, three years after Planck had received it.

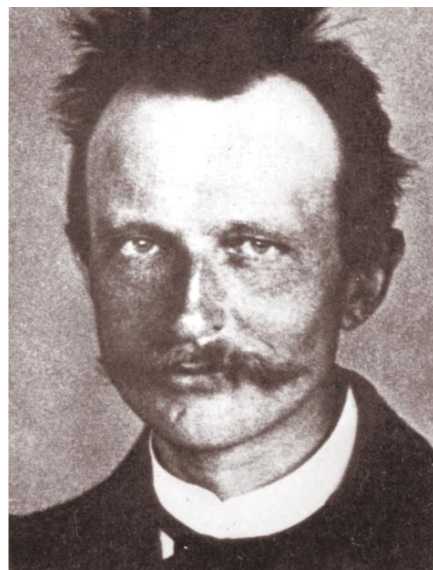
A change of perspective

It was also Einstein who first realized that the quantum hypothesis would lead to a major change in our view of the world, particularly by giving randomness a new and much more fundamental role than before. This discussion about interpreting quantum mechanics — which occupied the minds of many leaders in the field, especially after the formulation of modern quantum mechanics by Werner Heisenberg and Erwin Schrödinger in 1925–1926 — is still going to this day.

Heisenberg’s matrix mechanics of 1925 and Schrödinger’s wave mechanics of 1926, soon found to be equivalent theoretical descriptions of quantum phenomena, launched probably the most successful period of theoretical science in human history. In atomic physics, energy levels in atoms could at last be explained. And later, with the introduction of group theory into quantum mechanics, this explanation was extended even to complicated molecules. At the same time, simple molecules could be described quantitatively, and there were enormous successes in applying quantum mechanics to the solid state.

Hans Bethe, in his article “Quantum theory” in *More Things in Heaven and Earth*, celebrating the centennial of the American Physical Society in 1999, wrote: “1926, the year when I started graduate work, was a wonderful year for theoretical physicists. Whatever problem you tackled with the new

A new beginning: in 1900 Max Planck (right) articulated the concept that underpins quantum theory and that sparked debate between Einstein and Bohr at the 1927 Solvay congress (below).



tools of quantum mechanics could be successfully solved, and hundreds of problems, from the experimental work of many decades, were around, asking to be tackled."

Another important discovery in 1925 was Pauli's exclusion principle, which says that no two electrons can occupy the same quantum state. This principle plays a central role in many fields — for example in solid-state physics it helps explain the electrical conductivity of metals. It also tells us why chemical elements are so different. Modern quantum theory then entered a stage of maturity over the next two years, during which Paul Dirac developed the quantum theory of the electromagnetic field, the mother of all modern field theories, which are so important in the physics of elementary particles, and the relativistic quantum theory of the electron, which predicts the existence of antimatter.

The discussion of the philosophical interpretation of quantum mechanics continued in parallel with the enormous successes of the new theory. It culminated in the famous debate between Einstein and Niels Bohr, which began at the Solvay congresses of 1927 and 1930 and continued later in writing because of Einstein's emigration from Nazi Germany. Using a series of elegant *Gedanken* experiments — thought experiments — Einstein initially tried to show that quantum mechanics is inconsistent, in that it is possible to extract from an individual quantum phenomenon more information than is described by the limit represented by the Heisenberg uncertainty relation. Yet, in all instances, Bohr showed that consequent application of the new quantum laws avoided any contradictions.

Intellectual tug-of-war

Although Einstein clearly did not have the upper hand in his debate with Bohr, it is very much to his credit that he was one of the few who saw that quantum physics fundamentally challenges our view of the world. This is because it forces us to give up what can be called the naive classical realism so central not only to the views of physicists but also part of our approach to, and interpretation of, everyday life.

Finally, the famous Einstein–Podolsky–Rosen paper of 1935 introduced the idea of pairs of particles strongly correlated over large distances. Einstein had thought that these correlations would be a last-ditch stand of classical realism, because they hint at a possible classical model. But his hope to explain them using hidden properties of the individual systems was shown by John Bell in 1964 to lead to predictions that are in conflict with quantum mechanics. In fact, Schrödinger had realized in 1935 that such correlated systems are extremely non-classical. He coined the term 'entanglement' for these correlations and he called them "the essential characteristic" of quantum mechanics.



Here and not here: the Cheshire cat may have been the first observation of macroscopic superposition.

Technological progress since then has led to the possibility of performing not only many of the early *Gedanken* experiments but also a plethora of new ones. Most significantly, it is now possible to do real, detailed experiments with individual quantum systems such as individual photons, electrons, positrons, neutrons and atoms (even some made of antimatter) and molecules as large as fullerenes. Quantum experiments with more complicated systems have also become routine, where particles are entangled with each other just as originally postulated by Einstein, Boris Podolsky and Nathan Rosen. All modern experiments confirm the quantum predictions with unprecedented precision. And so, although one tiny loophole still remains for advocates of a classical world view, the evidence overwhelmingly suggests that a local realistic explanation of nature is not possible.

But the story does not stop here. As often in the history of physics, investigation of fundamentals has given rise to a new field. This fledgling field, which can be called the physics of quantum information, deals with the novel possibilities of encoding, transmitting and processing information through individual and entangled quantum systems. Quantum cryptography promises to provide us with a communication technology guaranteed to be secure against eavesdropping. Quantum teleportation can be seen as the possibility of directly transferring information from one system to another without that information first being read out and then transmitted by travelling on some carrier. And, as the most ambitious long-term goal, the quantum computer, if ever built, would give us novel computational techniques of unprecedented speed.

Although it is impossible to predict how a future quantum information technology will look, it is probably a safe bet that quantum laws will be directly relevant for communicating and processing information. This is because technological progress means that fewer and fewer electrons are needed to switch an individual bit in modern microprocessor chips — this is enshrined in Moore's law, which says that the number of

transistors on a chip doubles every 18 months. If this law continues to hold unabated, we will reach the quantum realm in about 20 years, when an individual bit is carried by just one electron.

The big increase in activity in fundamental experiments over the past two decades has also renewed the debate about the interpretation of the theory. Richard Feynman once commented, "I think I can safely say that nobody today understands quantum mechanics", and Sir Roger Penrose remarked that, although the theory agrees very well with all experiments and it is of profound mathematical beauty, it "makes absolutely no sense". So, where is the problem, if the theory fits all experimental results so nicely? The problem arises when we dare to ask what quantum mechanics might mean for our view of the world (*Weltanschauung*) in a broad sense. Can we safely, as many do, restrict the counterintuitive notions of quantum mechanics such as quantum superposition and entanglement to the microscopic world?

An absurd idea?

Schrödinger formulated his famous cat paradox in 1935, just to show how absurd the consequences of quantum mechanics are if applied to macroscopic or even living objects. We all know that a cat cannot be alive and dead at the same time as we never experience such a thing in everyday life.

Although the literature abounds with papers arguing that for one reason or another we may never expect to see superpositions of macroscopic objects, it might very well be that we should replace 'never' by 'not yet'. After all, who knows what experimental progress will bring in the next century, let alone the next millennium. And who knows what clever tricks our theoretical colleagues will find to circumvent the arguments currently put forward against the possibility of observing macroscopic superpositions. It is a safe bet that the research programme to demonstrate quantum phenomena for objects of increasing size and complexity will turn out to be one of the most interesting and

Evidence suggests that a local realistic explanation of nature is not possible.



hopefully fruitful avenues of research.

Even with the observation of macroscopic superposition still far away, many physicists are already puzzling over the so-called quantum measurement problem. This is the fact that although quantum mechanics makes perfectly valid predictions for statistical ensembles, in general it is not possible to make definite predictions for individual events, or, as John Bell put it, quantum mechanics does not explain “why events happen”. In the case of Schrödinger’s cat, even if we accept that the cat might be in a superposition state of ‘live’ and ‘dead’, there is no way to explain why, in a specific run of the experiments, we observe the specific outcome ‘alive’ or ‘dead’. We can only give their probabilities. This measurement problem has elicited many different responses from within the physics community.

One resolution to the problem suggested by some, for example Ghirardi, Rimini and Weber, and Pearle, modifies the evolution laws of quantum mechanics so that they become non-linear. But if measurements are precise enough, this will turn out to be in conflict with the predictions of standard quantum mechanics, which is linear. I venture that all future experiments will confirm the quantum predictions and thus result in conflict with nature for such non-linear theories.

Some interpretations try to define the problem away. These include the Many-Worlds interpretation, which assumes that all measurement results coexist in split universes. Another possible position is Bohm’s viewpoint, where particles actually have well-defined positions and momenta at all times but move in a quantum potential as a consequence of the Schrödinger equation. This latter point of view makes the same predictions as the standard theory and so does not teach us much that is new.

In the interpretive discussion, the early leaders in the development of quantum mechanics all had their individual points of view. For example, Dirac did not see such discussions as relevant, Schrödinger and Planck were realists, and others had rather unconventional points of view. To Heisenberg the roots of the problem were that the

Cartesian divide — the separation between *res cogitans* (that which thinks) and *res extensa* (that which is out there) — had deeply penetrated the human soul during the three centuries after Descartes. For Heisenberg, this was why the epistemological paradigm on which one could build the foundations of quantum mechanics had not been found yet. And Pauli even feels that here we have a recurrence of the medieval “*anima mundi*” (the soul of the world).

An informative theory

Notwithstanding the fact that there are a number of other interpretations, all with their own enthusiastic supporters, it seems that the measurement problem is trying to tell us something interesting.

One of the most careful thinkers on this matter, although not easily understood, is again Niels Bohr, who said: “There is no quantum world, there is only an abstract quantum physical description. It is wrong to think that the task of physics is to find out how nature is. Physics concerns what we can say about nature.”

Continuing Bohr’s line of thinking, it is suggestive to view quantum mechanics as a theory of information. Then the quantum state is just the representation of the information we have in a given situation. In general, the laws of quantum mechanics just permit us to make probabilistic predictions for possible future measurement results. In measurement, simply by observing an experimental result, we obtain novel information and so have to change the representation of our information, the quantum state. The fundamental laws themselves, such as the Schrödinger equation, are just expressions of conservation of information. But this makes a number of questions moot and just a matter of speculation. There are questions which make absolutely no sense — for example, through which of the two slits a particle passes when the double-slit interference experiment is observed, or whether Schrödinger’s cat really is alive or not.

It thus appears that after one century of quantum mechanics we have fascinating parallel developments. On the one hand we have new ways of processing information

using quantum mechanical laws. On the other hand we are gaining new insight into quantum mechanics itself by viewing it as an advanced theory of information.

One important open question is whether or not the laws of quantum mechanics will ever be found to play a significant role in biological systems, besides the obvious one of explaining their chemistry. In other words, have biological systems been selected by evolution to avoid the quantum domain so that they do not run into the problems of randomness, uncertainty and indefiniteness encountered there? Or are there at least some instances where biological systems can make use of the positive features such as quantum entanglement and quantum superposition? Clearly, this question is related to the one discussed above — is there an upper limit for the scale on which quantum phenomena can be shown to appear? In its most ambitious form this question asks whether the quantum might play any role in an understanding of mind.

Further pondering the future of quantum mechanics, it is probably safe to say that its basic rules are simple enough to be robust against change for a long time, if not for ever. Among these basic principles are the use of probability amplitudes whose squares give probabilities and which are superposed in a linear way, or the Pauli principle. Also, the mathematical beauty of the theory is a strong argument for its robustness.

At another frontier, the programme to quantify gravity (see accompanying article by Giovanni Amelino-Camelia on page 661), that aims to arrive at a description of space and time consistent with quantum mechanics, for some physicists points to the necessity for a deep revision of our foundations.

If anything is certain then it is the fact that physics will be completely different 100 years from now. Suggestions made today by some that we are close to an end of science or close to finding the final theory will be as much demonstrations of the limitations of human imagination as similar suggestions made in the nineteenth century. Then, for example, Max Planck was told in 1874 by the Munich physicist Philipp von Jolly to study something else, as all fundamental laws were known and all that was left to physicists was to fill in a few remaining details. ■

Anton Zeilinger is in the Institut für Experimentalphysik, Universität Wien, Boltzmanngasse 5, 1090 Wien, Austria.

Further reading

1. Bouwmeester, D., Ekert, A. & Zeilinger A. (eds) *The Physics of Quantum Information* (Springer, Berlin, 2000).
2. Bederson, B. (ed.) *More Things in Heaven and Earth: A Celebration of Physics at the Millennium* (Springer, New York, 1999).
3. Amelino-Camelia, G. *Nature* **408**, 661–664 (2000).
4. Schilpp, P. A. (ed.) *Albert Einstein: Philosopher-Scientist* (Open Court, La Salle, 1949).
5. Jammer, M. *The Philosophy of Quantum Mechanics* (Wiley, New York, 1974).
6. Wheeler, J. A. & Zurek, W. H., *Quantum Theory and Measurement* (Princeton Univ. Press, 1983).

All human life is here

A tour of anthropology from bipedalism to environmentalism.



Human Natures: Genes, Cultures, and the Human Prospect

by Paul R. Ehrlich

Island Press: 2000. 576 pp. \$29.95

Vaclav Smil

This is a new Paul Ehrlich, relaxed, reminiscing, nonprescriptive and unexpectedly upbeat. Ever since his book *The Population Bomb* (Ballantine, 1968), Ehrlich has conducted a crusade aimed at not just stopping the growth of the global population but even reducing its existing numbers. These prescriptions have been based on his conviction that we have already exceeded some of the planet's productive limits, and this conclusion has led him to publish repeated prophecies of the near-imminent demise of humanity and the unravelling of the biosphere. Not so this time: Ehrlich is largely content just to observe and to explain as he gives us his systematic introduction to evolutionary anthropology.

Along the way he shares personal anecdotes, preferences, foibles and weaknesses. And so we are taken back to the moment when he captured his first Arctic butterfly (1952) and when he was repeatedly bitten by water snakes in Lake Erie (1957). We learn that he gets backache, is lousy at mathematics and language learning, must be corrected by his granddaughter when pronouncing chocolate in Spanish — and that he craves rare beef and has some expensive tastes (Chateau Mouton 1945, whose smell he is able to recall instantly).

But these asides just leaven a serious story as Ehrlich delivers a systematic and determinedly thorough review of the multifaceted subject of human evolution. There is the unmistakable impression of a committed scientist visiting a new field and trying hard not to leave out any of the iconic facts, terms and images. Consequently, Ehrlich has given

us a readable mini-encyclopedia of modern evolutionary anthropology, and not too far into it I became caught up in the guessing game: will he include this or that expected fact? Almost invariably he does. His quest for inclusion and detailed explanation is also reflected by the allocation of space: the text takes up only about 60%, notes (often expansive, enlivened by reminiscences and personal asides and definitely worth reading), set in a small type, add 100 pages, or almost a quarter of the entire length, and references and index take another 100 pages.

And so we encounter a multitude of topics all too well known to undergraduates in physical anthropology courses: there is the rise of bipedalism, the migration out of Africa with its Olduvai gorge, Lucy and the Turkana boy, and the sad fate of the Neanderthals. And, of course, there are discussions of natural selection, human physical variability, *Drosophila* mutations, and sexually insatiable bonobos. But, as the book's subtitle makes clear, Ehrlich is much more interested in mental and cultural evolution and, once again, he gathers in every pertinent, oft-repeated fact and conclusion that has graced the pages of the anthropological, psychological and sociological literature of the past generation.

This includes a discussion of IQ scores and *The Bell Curve* (a controversial book by Richard J. Herrnstein and Charles Murray that sees human intelligence in a genetically deterministic way); affectionate and aggressive Gombe chimpanzees; the ape Kanzi's ability to understand commands and Noam Chomsky's idea that a universal grammar is in a child's mind at birth; belief in God; and the power of chance in determining human future (on that score he ascribes excessive consequences to the Battle of Midway).

Many of the illustrations have also seen too many reprints: *Archaeopteryx*; a map

Seen it all before: Easter Island's statues pay mute testimony to the demise of their builders.

of the functional areas of the cortex with its grotesquely shaped faces and limbs; ambiguous-perception figures (vase-face, the Necker cube); brain-damage survivor Phineas Gage's skull penetrated by the tamping iron; cross-sections of chimpanzee and human heads showing their tongues and larynxes. And, yet again, Ehrlich's iconic quest continues when he moves onto matters of cultural evolution: there are vicious Yanomami, destructive Manganians and virtuous Tikopians, the moai statues of Easter Island bearing a mute testimony to the demise of their builders...

Anthropology majors will enjoy the book, as it goes over so much of their familiar ground — but it does so in a refreshing way: reinforcement through interesting recapitulation. And even those research anthropologists who will find little in the book that is fundamentally new will, I would think, enjoy its scope, arguments and insights and appreciate its assembly of notes and references. Readers from outside the book's main disciplines will get the greatest reward, as they will learn a great deal in a commendably interlinked fashion.

But the book's greatest surprise comes on its penultimate page, where Ehrlich joins the ranks of the realists. He concludes that our challenge is how to deal sensibly with both nature (learning to be environmentalists for the sake of the biosphere) and our natures (learning how to forge a better society). Who could disagree? And then — after admitting that he tends "to be optimistic in thinking that we *can* do it but pessimistic about whether we *will* do it" — he writes that "maybe even the latter pessimism is unjustified".

For many years Ehrlich's writings represented an excessive pessimism, and they were often contrasted with the late free-

market economist Julian Simon's sometimes overenthusiastic optimism that the human intellect was the Earth's only important resource. Neither attitude advances a serious discourse of complex scientific matters whose understanding helps to shape our fate. For many years, when reading Ehrlich, I wished that his incisive conclusions and passionate advocacies based on his wide-ranging understanding were not couched in such catastrophic tones. That is why I welcome this book: it will not be a revelation to anyone reasonably well versed in modern evolutionary anthropology — but it is a spirited, valuable and enjoyable contribution to an interdisciplinary understanding of our complex natures. ■

Vaclav Smil is at the University of Manitoba, Winnipeg R3T 2N2, Canada.

An uncertain principal

Quantum Reflections

edited by John Ellis & Daniele Amati
Cambridge University Press: 2000. 202 pp.
£30, \$49.95

Daniel Greenberger

John Bell has been dead for ten years now, but his influence on quantum theory is greater than ever. This is a book dedicated to his memory. To those in the small field devoted to the fundamental problems of interpreting quantum theory, he was a colossal figure. Had he not died prematurely, he would almost certainly have won a Nobel prize. After all, how many physicists have actually started a new field of research?

When I was a graduate student back in the late fifties, it was considered bad form to worry about the foundations of quantum theory. To do so in any serious way was a guaranteed road to destroying one's career. We were taught that the founding fathers had said all there was on this matter in the thirties, and that there was nothing useful to add to the subject.

There had been Einstein's objections to the theory, but they had been laid to rest by Niels Bohr. Only his EPR objections (published in a 1935 paper by Einstein, Boris Podolsky and Nathan Rosen) had not been definitively answered, but they were not accessible to experiment. The idea of constructing an interpretation of the theory that involved hidden variables had been proven impossible by John von Neumann, and that was that. (David Bohm had actually constructed such a theory in the fifties, but it was not well known, and certainly not taken seriously by most of the people who did know about it.)

In the sixties Bell demolished von Neu-

mann's 'proof', and showed that it was experimentally possible to distinguish between the opposing positions of Einstein and Bohr. He did this by deriving the first of the now-famous Bell inequalities — any local hidden-variable theory built along the lines envisaged by EPR would lead to experimental results that satisfy this inequality, but quantum theory leads to results that violate it.

It was this that gave rise to the field of testing quantum theory against interpretations using hidden variables — tests that quantum theory has passed decisively. This has led to increasingly sophisticated experiments, and the realization of the importance of 'entangled states' in delineating the non-classical behaviour of quantum states. This has played a major role in the development of the fields of quantum cryptography and quantum computing, and the gradual absorption of all these fields into the mainstream of quantum mechanics, both on the theoretical and experimental side.

Nonetheless, Bell was very circumspect about his work on these fundamentals, and treated it as a sort of hobby. Few of his colleagues at the European particle-physics laboratory CERN were aware of it, or of the fact that there was a community of their colleagues to whom Bell was a hero. And so he led several disparate lives in physics. Yet everyone who knew him saw that he was special — a very careful and deep thinker about everything, a passionate partisan for whatever he believed in, but a man of quick wit with a twinkle in his eye. And although he would not have liked it, he is becoming a cult figure in the physics community — hence this book.

There is a short biographical note by his wife, Mary Bell, who is also a physicist, and an

introduction by the editors. This is followed by nine essays, most of them describing areas in which Bell was interested, and a few of which are also sort of reminiscences. By and large, they are of a high quality, written for the physicist who is not a specialist in Bell's fields. There are fine chapters by Alain Aspect and Abner Shimony, two first-class physicists and expositors, on the experimental and theoretical aspects, respectively, of the Bell inequalities. Roger Penrose has written a more technical piece which demonstrates his uncanny geometrical intuition in higher dimensions (Penrose is the only person I know who would feel at home in an Escher drawing).

Helmut Rauch contributes a beautiful article on non-classical aspects of neutron interferometry. The neutron interferometer was the first instrument that allowed one to actually perform a series of what were formerly gedanken experiments in quantum theory. (Gedanken experiments are thought experiments that are too hard to do in a real laboratory. But now many, such as rotating a neutron by 360° in order to show that the wave function reverses sign, have actually been done.) And GianCarlo Ghirardi explains the way that he and his colleagues implemented a nonlinear extension of the Schrödinger equation that would actually cause the physical collapse of a wave packet for a macroscopic state.

It might be thought that because experiments on Bell's inequalities have gone a long way towards justifying quantum theory, Bell himself would have been a staunch supporter of the theory. But that is far from the case. He thought that although Einstein's ideas were thereby disproved, Einstein's reasoning was far better than Bohr's, and that Einstein had been defeated by cruel mother nature,



Entangled: Bell, here with his physicist wife Mary, was a sceptical investigator of quantum theory.

not poor logic. Bell thought that the conventional ideas about measurement theory, with its 'collapsing' wave packets, were fundamentally flawed, and he was very sympathetic towards ideas for improving the situation, such as the nonlinear collapse theory. Although he readily conceded that the standard theory was fine "for all practical purposes" (which he derided as FAPP), he insisted that measurements had to be described by the equations themselves, and not as something postulated independently.

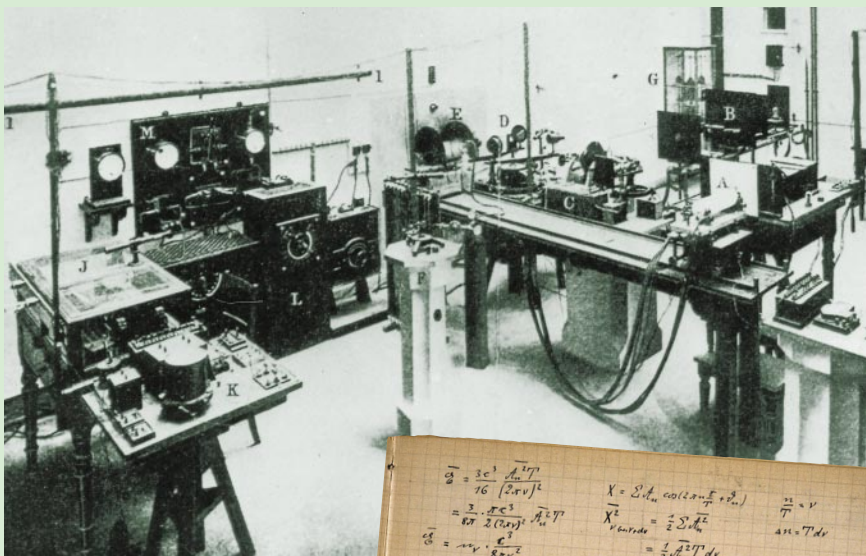
On the other side of this argument, Kurt Gottfried defends the standard interpretation of the theory, although he cedes many points to Bell's attack. This argument may seem convincing to most working physicists, but Bell would probably have replied with his wonderfully dismissive word: FAPP-TRAP!

Finally, there are three articles by collaborators of Bell on other topics. One of these is a fascinating article by Jon Magne Leinaas on the Unruh effect, as seen in the behaviour of electrons in an accelerator (according to this amazing effect, a particle in an accelerated system 'sees' the transformed vacuum as black-body radiation at a finite temperature.)

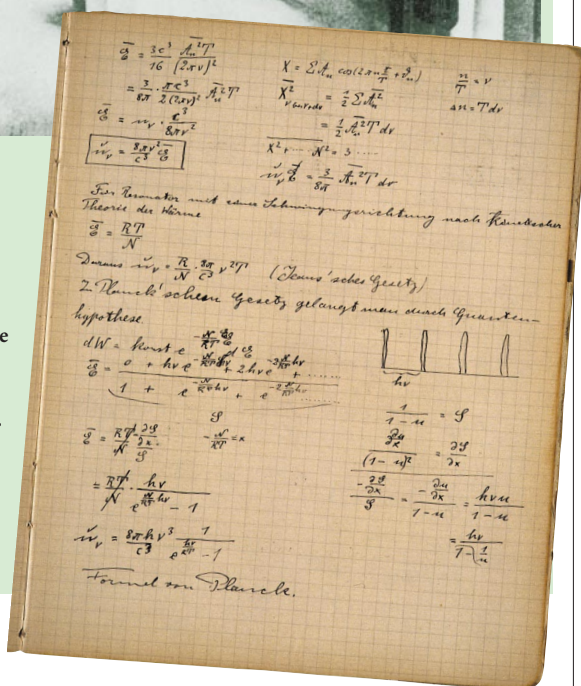
The editors have managed to put together a book honouring John Bell written by people who are not only world-class physicists themselves, but who are all people one would refer to as *Menschen* — people of great humanity. This makes the book an especially fitting tribute to a man whose humanity was an intrinsic part of his greatness.

Daniel Greenberger is in the Department of Physics, City College of New York, New York, New York 10031, USA.

Relics of the quantum century



Until 15 February 2001, visitors to Berlin's Staatsbibliothek am Kulturforum can see an exhibition of artefacts connected with quantum physics. These include a reconstruction of Max Planck's laboratory of 1895 (above), where the germ of quantum theory took root, and a page from Einstein's 'Zurich notebook', dating from around 1911. Einstein was trying to convince his colleagues that Planck's law was a break from classical physics. Schrödinger's cat was unavailable for public appearances.



Observing the astronomer

Henry Norris Russell: Dean of American Astronomers

by David H. DeVorkin

Princeton University Press: 2000. 499 pp. £30, \$49.50

Dorrit Hoffleit

More than any other American astronomer, Henry Norris Russell (1877–1957) believed that observational astronomy was only of use if it was directed specifically at solving problems concerned with the constitution and evolution of the stars. For, before 1920, so little was known about atomic theory that progress in interpreting stellar spectra was extremely erratic. David DeVorkin has carried out a remarkably thorough search into both Russell's family life and his scientific accomplishments.

Russell's father was a Presbyterian minister, and Henry grew up with strong religious principles. However, when he married Lucy

Cole, an Episcopalian, the couple attended the Presbyterian and the Episcopal churches on alternate Sundays. They had four children, three girls and a boy. The two eldest were twin girls, of whom one was seriously handicapped with cerebral palsy. Whenever his wife fell ill, Russell was greatly concerned lest he should be left to look after their small children alone. He himself was always in frail health.

Russell's nervously active mind sometimes led him astray in his ideas about stellar composition and evolution, but his conclusions were always reasonable given the knowledge of the laws of physics at the time. This biography shows how theories once considered logical may later be discredited when newer laws of physics are discovered.

Russell was the most brilliant student under the astronomer Charles Young at Princeton. For his PhD he studied the light curves of Algol-type eclipsing binaries — two stars revolving around one another — as a means of determining the masses of the component stars. Only by using the gravitational effects of the stars on one another could their masses be determined; the

masses of single stars could not be worked out.

After he had obtained his PhD, Russell spent some time with Arthur Hinks in Cambridge, England, determining the parallaxes, or distances, of stars. From these and the apparent magnitudes, or relative brightnesses, of the stars, he determined their absolute magnitudes — how bright they would look if all were at the same standard distance. This work confirmed and improved upon the earlier important discovery by the Danish astronomer Ejnar Hertzsprung that stars of the same colour, or spectral class, do not all have the same intrinsic luminosity. From these results Russell developed an attractive theory for the evolution of stars that was accepted for many years. It was finally abandoned when, after more was known about the properties of atoms, it was found that the stars Russell assumed to be the oldest turned out to be the youngest.

Russell was an excellent adviser to graduate students, but he had no great interest in teaching undergraduate astronomy. As



Top of the heap: Russell was a domineering personality whose "word was law" in US astronomy.

director of the Princeton Observatory, he delegated undergraduate teaching to two able assistants, Raymond Dugan and John Stewart, who also collaborated with him on a revision of Young's manual, *Astronomy*, the first revisions to which were published in 1926–27, and final revisions in 1945.

In 1900 Russell was asked to write monthly articles for *Scientific American*, a task he enjoyed for 42 years. These articles were written with the scientifically minded layperson in mind, but otherwise, Russell paid little attention to popularizing astronomy. In fact, he disapproved of the considerable time Harlow Shapley spent on public education during his directorship of the Harvard Observatory, even though such efforts helped in the necessary fund-raising. Russell also criticized Shapley — his former best graduate student at Princeton — for continuing his predecessor Edward Pickering's "factory" type observational programmes. These programmes were aimed at producing photometric and spectroscopic catalogues, whereas Russell believed observational programmes should be directed at solving specific theoretical problems in astrophysics.

In testing his own theories, Russell was not interested in making telescopic observations himself. Instead, he got permission to examine the collection of spectroscopic plates at the Harvard College Observatory.

US astronomer George Hale had consulted Russell about suitable programmes for the telescopes he built at the Mount Wilson Observatory above the Los Angeles basin. Russell was consequently a welcome visitor at Mount Wilson, and soon persuaded the Princeton administration to give him an annual three months leave of absence to spend there. He also made frequent visits to Harvard. Indeed, Russell seems to have been attempting to direct the research at all three observatories. At Harvard he tried to steer

Shapley into concentrating on the fields of astrophysics with which he himself was most concerned. Although Shapley obviously felt that in any controversy over the composition or evolution of stars, his mentor Russell could not be wrong, he still pursued the topics that most interested himself — the structure of the Milky Way and the distributions in space of external galaxies — subjects to which Russell paid little heed.

The first graduate student to obtain a PhD in astronomy at Harvard or Radcliffe was the British woman Cecilia Payne. In her 1925 thesis she correctly identified hydrogen as the most prevalent atom in the stars. Russell objected that her conclusion could not be correct because he had worked out that the composition of the stars, especially that of the Sun, was the same as the more metallic composition of the Earth's crust. Payne was forced to water down her thesis by saying that there must be a mistake in her analysis because her conclusion differed from then prevalent beliefs. She said of Russell: "His word was law. If a piece of work received his imprimatur, it could be published; if not, it must be set aside and its author had a hard row to hoe." In 1929, from an examination of Mount Wilson spectra, Russell himself reached Payne's conclusion about the prevalence of hydrogen. And in 1934, when Russell was considering whom he should be training as his successor on his retirement in 1943, he lamented that the best-qualified person, "alas, is a woman" — meaning Payne.

After Russell's son-in-law, Frank Edmondson, obtained his PhD at Harvard, Russell got him an offer of a research position at the University of Virginia. Edmondson declined, preferring a position at his undergraduate University of Indiana, where he would be in line for the directorship of the observatory. Russell strongly disapproved, stating that administration is anathema to

good research. But, like Shapley, Edmondson followed his own inclination, and returned to Indiana.

In his final chapter, "Russell's Universe", DeVorkin comments that Russell "helped transform American astrophysics into a wholly physical discipline. Establishing this framework, more than any one discovery or application that bears his name, is Russell's greatest legacy to astronomy." Yet DeVorkin continues, "much of his work was not really his own, not new ... he typically seized upon the incomplete work of others and tried to confirm or refute it".

This comprehensive biography should be of interest not only to historians of science and students of astronomy, but also to psychologists who might enjoy analysing this brilliant, domineering personality. ■

Dorrit Hoffleit is in the Department of Astronomy, Yale University, 260 Whitney Avenue, New Haven, Connecticut 06520-8101, USA.

New in paperback

Vivisection or Science? An Investigation into Testing Drugs & Safeguarding Health

by Pietro Croce *Zed Books, £14.95*

Energies: An Illustrated Guide to Biosphere and Civilization

by Vaclav Smil *MIT, \$15.95, £9.95*

In the Eye of the Beholder: The Science of Face Perception

by Vicki Bruce & Andy Young
Oxford University Press, £14.99
"Explains the importance of the face, how we extract the information it contains, and reveal what it all means." *Nature* 393, 752 (1998)

Visual Intelligence: How We Create What We See

by Donald D. Hoffman
W. W. Norton, \$17.95, £12.95
"I would be surprised if even the most experienced visual scientists do not learn much from Hoffman's guidance ... The moral of *Visual Intelligence* is that we have spent so long in Plato's cave that our brains can now derive what is really out there from the merest flickering of shadows on the wall." John C. Marshall *Nature* 396, 36 (1998)

Jacobson's Organ: And the Remarkable Nature of Smell

by Lyall Watson *Penguin, £5.99*
"Lyall Watson's easy, flowing text picks its way through the comparative anatomy, biochemistry and evolutionary biology of olfactory communication ... Its uncritical approach will only fuel further speculation about the supposed supernatural powers of the sense of smell." Michael Stoddart, *Nature* 405, 124 (2000)

A chair of one's own

The upper reaches of academe remain stubbornly inaccessible to women.

Christine Wennerås and Agnes Wold

If a woman is to write fiction, said Virginia Woolf, she will need money and a room of her own. Likewise, if a woman is to do science she will need grants and a laboratory of her own. The female scientist will also strive for a chair of her own, but she will find it elusive. Although women hold over half of the bachelor's degrees in Europe, they hold just one-tenth of full professorships. Despite decades of debate and measures directed towards making the top levels of academe accessible to women, they remain stubbornly chairless.

Wherever they are, female academics tread a harsher pathway than their male colleagues. US female medical-school graduates are more likely than their male classmates to pursue academic careers, but they are less than half as likely to be promoted to professors. In Italy, it is twice as hard for female senior researchers supported by the National Research Council to become research directors compared with their male counterparts.

In countries where the proportion of women among the professorate is even lower than in the United States and Italy, the hurdles facing women academics are even higher. In Germany, 25% of professors would have been female, instead of the 4%

seen today, if female university graduates had been able to follow male career paths. If Prometheus had lived today, he would probably have been a female scientist.

Family and children are often blamed for women's poor academic success, but studies refute this explanation. In the United States, Finland and Norway, female researchers with children are actually more productive than their childless female colleagues. The true reason for women scientists' sluggish careers must be sought within academia itself.

During the millennium of their existence, universities have devised more or less ingenious strategies to exclude womankind. The coarsest schemes prohibited women from entering the university and attending lectures, often with the backing of legislation. A more refined line of conduct was to allow women to study, but with severe limitations. For example, only certain disciplines were open to them. Women were also frequently denied the right to take degrees, and — as Woolf bitterly experienced — access to university libraries was carefully circumscribed for women scholars. Today, women academics don't face such formidable opposition, yet still they lag behind. Why?

Talent alone does not determine a scientist's career. Time, space and money must be added to the brew. But nowhere in the world

are these shared equally between the sexes. In the United Kingdom, only 20% of Medical Research Council or Wellcome Trust grants end up in the pockets of female researchers, who make up 44% of the biomedical academic staff. At the US National Cancer Institute, women researchers on average receive less than two-thirds of the budget and 63% of the research staff compared with male peers of equal seniority. This fact alone can account for the apparent lower scientific productivity of these female scientists.

Identical pieces of work, for example paintings or essays, are often judged more severely if they are assumed to be made by a woman. Scientists are not exempt from the prejudices against women that prevail to this day in all societies. Three years ago, we examined the peer-review process at the Swedish Medical Research Council and found that women had to produce twice as many scientific papers of equivalent quality to those written by men to be considered equally competent. The systematic underestimation of female performance is particularly deleterious in fields such as science, where individuals are constantly evaluated. Repeated small injustices accumulate to produce visible differences in career paths between the sexes. Only if she has excellent contacts can a woman compete on equal terms with a man.

Women's slower pace of rank advancement in itself hampers their scientific productivity. High academic rank makes it more likely that people will include you on their author lists. A junior scientist can produce one good paper per year, a leader of a small research group three to five, whereas the principal investigator of a large team can easily churn out 20. This creates a vicious circle, in which low rank feeds feeble productivity, succeeded by poor career advancement. To those who have, more will be given.

Junior scientists' frustration at the pace of their scientific productivity is normal at the beginning of their careers, when they do most of the benchwork by themselves. But female scientists tend to remain at this level their entire working lives. One should thus not underestimate the importance of having a chair of one's own. To return to Virginia Woolf: "Nobody in their senses could fail to detect the dominance of the professor. His was the power and the money and the influence." It is high time for female scientists to become women of influence. ■

Christine Wennerås and Agnes Wold are associate professors in the departments of clinical bacteriology and clinical immunology, respectively, at the Sahlgrenska University Hospital, Göteborg, Sweden.



Personal space: Marie Curie was a rarity among women scientists in having a lab at her disposal.

BETTMANN/CORBIS

Murphy's cat

Take care who sits next to you at conferences...

Joan D. Vinge

The scene: a lecture hall, slowly filling. MURPHY, carrying a canvas bag, sits down beside the distinguished researcher waiting to speak.

MURPHY. So, you got the copy of *Nature's Retrospective on the Future* in your e-load, too. Amusing, huh?

PROFESSOR [*dubiously*]. Everyone got it, at registration. And while I found it fascinating, 'amusing' never occurred to me.

MURPHY. Well, considering your research, I suppose not. ... By the way, not everyone got it: there was a glitch in the download. Some people got *The Tao of Pooh*.

PROFESSOR. Do I know you? Have we ever met?

MURPHY. Sure. I'm Murphy; we bump into each other all the time. Literally. You really need to do something about that tunnel vision of yours. [*Suddenly addresses the bag, which has begun to flop around on the floor*] Settle down!

PROFESSOR [*Rising to leave, sits back down*]. It's alive—?

MURPHY. And always at the most inopportune times. [*Picks up the sack, peers into it*] It's only sixteen hundred hours, Dingy; dinner's at eighteen. Take a nap, baby.

PROFESSOR. Good Lord—

MURPHY. No, just my cat. Schrödinger gave him to me.

PROFESSOR. Schrödinger's cat? This is absurd; there was no actual cat. It was all hypothetical.

MURPHY [*Holds up the sack*]. Tell that to the cat.

PROFESSOR [*Looks in, aghast*]. It's ... dead.

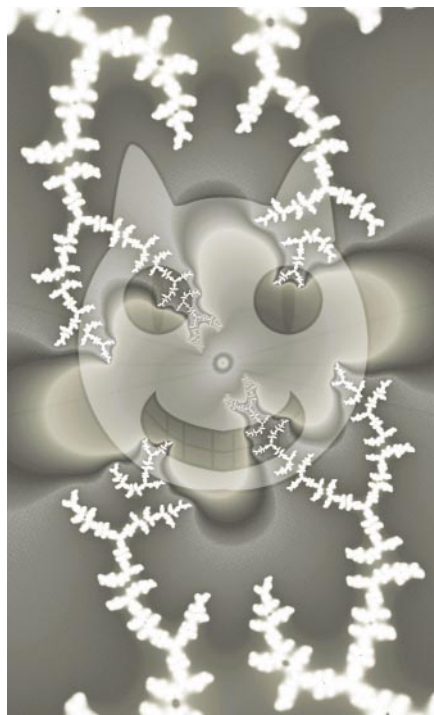
MURPHY. You miss my point. But he never misses dinner. A free vacation is the only reason we attend these snot-fests anymore. No one comes to my presentations, and no one believes he exists. He's been dead and alive for nearly a century, after all. Being a cat, he's sensitive to snubs.

PROFESSOR. And you two ... converse, do you?

MURPHY. This isn't *Waiting for Godot*. Don't be ridiculous. Our telepathic affinity is extremely high. As Sheldrake said, back in the last century, even humans possess the largely untapped potential—

PROFESSOR. Rupert Sheldrake? Oh please. Leave 'morphic resonance' in the dustbin where it belongs. No experiment ever found his 'mystery force field'.

MURPHY. But that was the point, accord-



ing to David Bohm. Chopping up dead things to discover the 'secret of life' is absurd.

PROFESSOR. I do not 'chop up dead things'. I work with nanotechnology. Besides, there is no evidence that we need to go beyond the biochemical to explain life. That is all the 'mystery' you need. And any day now we will have unravelled it completely—

MURPHY. Promises, promises ... [*Touches the Professor's shoulder with a fingertip; a spark of static leaps between them*] Boy, I hate these synthetic fabrics; I keep having to ground myself.

PROFESSOR. In reality, no doubt.

MURPHY. Cheap shot; bad aim. Speaking of reality, it was a real shame about those 'clone-the-frozen-head' experiments. Yuck.

PROFESSOR. Where did you hear about that?

MURPHY. I was there. Although as usual no one realized it. But about bioelectrics, Professor: where does the energy go, when something dies? Of course my cat understands all that better than I do. Not my department, as they say.

PROFESSOR. Obviously. Just what is your 'department', by the way?

MURPHY. Changing the subject. What do you think of Dürr's work, when he was at the Planck Institute? Or Lukens and Friedman's Y2K article from *Nature*, showing that the macrocosmic entrainment of microcosmic forces actually exists? It's in the

Retrospective packet. For so long we sought the submolecular soul of our 'orderly' existence, only to discover that it's utter chaos! Is that humour on a cosmic scale, or what?

PROFESSOR. Are you on drugs?

MURPHY. I get high on life, that's cosmic enough. But as I was saying: even better, the *Retrospective* has these odd little essays predicting 'the Future'—basically our present. They all seem to predict the perfect triumph of artificial intelligence over bestial human nature. Acrawl with nanobots, we have transformed ourselves into our own successors—or else humans are obsolete, out to pasture on a planet saved from ecodisaster by the wisdom of AIs. And nothing goes wrong along the way. As if—!

PROFESSOR. Murphy—

MURPHY. They don't even consider the socioeconomic aspects, let alone selfishness, phobias, or survival instinct! Articles written by robots about robots in a magazine called *Nature*!

PROFESSOR. Murphy, everything—including my patience—has its limits. Controlling nature, making every atom in it dance to my tune, is what I live to achieve. I'm certain that within the next 30 years—

MURPHY. Yadda, yadda. How about those AIs that live in cyberspace, that started out as minuscule bits of program? But then they linked up, rewrote their programming, mutated and grew, until all we can do now is try not to tick them off. Have you ever had one of Tildon's biomorphic 'bots malfunction and eat your shorts? How can you be sure a billion nanobots in your blood won't suddenly decide to turn you into a giant tumour?

PROFESSOR. I will be addressing control measures in my—

MURPHY [*Rises, picks up the sack*]. Well, gotta go.

PROFESSOR. You're leaving? Before my presentation?

MURPHY. I've heard it. And I have an online rendezvous with my sweet patootie before dinner. But my feminine intuition tells me we'll run into each other again. Soon.

PROFESSOR. What—?

MURPHY. I never said I wasn't a woman.

PROFESSOR. Wait—at least tell me your field of study?

MURPHY. Chaos theory. I have a law named after me, in fact. Naturally, it's the only law in the universe that always functions.

PROFESSOR. Which is—?

MURPHY. Take a wild guess. I dare you. ■

Joan D. Vinge has won two Hugo Awards for her science fiction. Her most recent novel is *Tangled Up In Blue*.

What drives climate?

Lee R. Kump

Variation in the amount of CO₂ in the atmosphere is usually taken to be the main cause of climate change on geological timescales. The apparent exceptions to the link threaten to undermine that view.

In a worst-case scenario, fossil-fuel burning will pump up concentrations of CO₂ in the atmosphere to levels higher than those of the past several tens of millions of years^{1,2}. Hence the intense research on the long-term behaviour of the carbon cycle. The aim is to understand the relationship between changes in atmospheric concentrations of CO₂ and climate swings from 'icehouse' (presence of large continental ice sheets) to 'greenhouse' (ice-free) conditions over the Phanerozoic eon — the past 544 million years of Earth's history. Various geochemical measurements and numerical models of the carbon cycle have been used to reconstruct ancient CO₂ levels³; such measurements are known as proxies because they are indirect, in this case being based on the carbon isotopic composition of soil minerals or organic matter in marine sediments. Climate reconstructions, of temperature in particular, have largely been derived from a rather incomplete record from fossil and geological indicators³.

On page 698 of this issue, Veizer *et al.*⁴ present an exquisite temperature record that is based on the oxygen isotopic composition of tropical marine fossils and that broadly mimics the inferred oscillations of global climate on 100-million-year timescales. However, calculations using proxy data for atmospheric CO₂ reveal mismatches with both the climate and the oxygen isotope records. Veizer *et al.* thus cast doubt on the prevailing view that most, if not all, of the climate swings in the Phanerozoic have been linked to variations in atmospheric CO₂.

There are two primary stable isotopes of oxygen, the more abundant ¹⁶O and the less abundant ¹⁸O. Isotopic compositions of various materials are reported as d¹⁸O, a standardized ratio of ¹⁸O/¹⁶O. Marine fossils acquired their oxygen isotopic composition when the organism was alive and accumulating calcium carbonate or silica from sea water to make its shell or skeleton. The d¹⁸O of this material reflects both the isotopic composition of the sea water in which the organisms lived, and the extent of preferential uptake of the isotopes during calcification or silicification; that uptake depended on the water temperature and other physiological factors that can be species-specific. During and after fossilization, however, fluids in sediment pores can shift oxygen isotopic compositions away from the original value, potentially compromising the

fossil's value as an indicator of temperature.

Veizer *et al.*⁴ have been very selective in their sample selection and analysis, using only those fossils that pass several criteria of preservation⁵. Nevertheless, they observe a long-term (100-million-year or more) trend of decreasing d¹⁸O with sample age that many would argue reflects the greater likelihood that the older samples have undergone post-depositional alteration. Veizer *et al.* discount that conclusion. They 'de-trend' the data, though, because they feel that the long-term trend is caused by tectonic processes, not climate. This then allows them to concentrate on the oscillations of d¹⁸O that coincide quite nicely with climate swings during the Phanerozoic, as inferred from indicators such as the occurrence and distribution of glacial deposits and climate-diagnostic fossil assemblages. This correspondence is encouraging, because it justifies the use of the isotope record to fill in the gaps in the other, relatively discontinuous climate records. Taken at face value, the isotopic swings imply

that the temperature of the sea surface in the tropics (the habitat of most of the fossilized organisms studied) has varied by as much as 13 °C over the Phanerozoic.

Veizer *et al.* acknowledge that part of the record must reflect changes in ice-sheet size. That is, the d¹⁸O of the ocean increases as ice sheets preferentially extract H₂O with ¹⁶O from the oceans as they grow, which increases the d¹⁸O of the shell of the organism independently of any temperature effects on d¹⁸O. Taking the ice-volume effects into consideration, they conclude that sea surface temperatures in the tropics varied by about 9 °C in the Phanerozoic. The inference here is that climate change has been global, rather than being restricted to high latitudes where the direct effects of ice-sheet expansion and retraction are greatest.

Not content with this result, important though it is, the authors went further. They proceeded to address the question of whether the variations in atmospheric CO₂ inferred from carbon isotopic proxies are

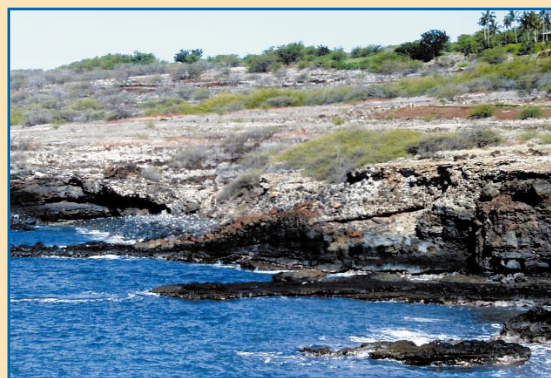
Earth science

Not making waves

When Bill Gates got married on the Hawaiian island of Lana'i, he might have wondered how the coral-rich gravel outcrops shown here came to be 30 metres above sea level. If so, he would have been told that these gravels, and others even higher up, were deposited by a giant tsunami 105,000 years ago.

The tsunami hypothesis came from the chaotic nature of the gravels, and the fact that other nearby Hawaiian islands are subsiding — so that deposits formed at sea level 100,000 years ago would not be expected to be at high elevations today. But the tsunami would have had to be truly 'giant', with a run-up of hundreds of metres, compared with a modern-day Hawaiian record of only 17 metres.

Elsewhere in this issue (*Nature* 408, 675–681; 2000),



Ken Rubin and colleagues describe a test of the tsunami hypothesis. They find that the gravels contain corals of two distinct ages (135,000 and 240,000 years old), both from times when sea level was high. From this and detailed mapping studies, they conclude that the gravels were in fact deposited not in a single catastrophic event, but by typical Hawaiian

coastal processes such as those forming the cobble beach towards the left of the photo.

And what of the subsidence problem? It turns out that recent flexure models of the Pacific plate can explain concurrent subsidence of some islands and uplift on others, including Lana'i. Uplifting experiences for both the island and the computer magnate, then. **Laura Garwin**

consistent with the climate history from oxygen isotopes and the general notion that CO₂ is a fundamental driver of climate. They use an energy-balance climate model incorporating the proxy record of atmospheric CO₂ to determine tropical sea surface temperatures during the Phanerozoic. The temperatures predicted with this model bear little resemblance to the climate record. Either the CO₂ proxies are flawed, or our understanding of the relationship between CO₂ and climate (as expressed in the climate model) needs rethinking. From the title of their article, "Evidence for decoupling of atmospheric CO₂ and global climate during the Phanerozoic eon", the authors evidently feel that CO₂ may not be the climate driver it has been made out to be.

Such a conclusion deserves close scrutiny, because the policy implications are huge. The geological record is our best hope of establishing a correspondence between atmospheric levels of CO₂ and climate, and understanding the likely consequences of fossil-fuel burning. If large changes in atmospheric CO₂ in the past have not produced the climate response we thought they had, that undermines the case for reducing fossil-fuel emissions.

The authors concentrate on two mismatches: the glaciation of the Late Ordovician period (440 million years ago) and the cool climate of the Jurassic and early Cretaceous (approximately 220–120 million years ago), both of which coincided with proxy indications of high levels of atmospheric CO₂. These seeming paradoxes are not new, however, and the Late Ordovician mismatch has been particularly well studied^{6,7}. Solar physicists argue that the Sun was then some 5% less luminous than it is today, so higher atmospheric CO₂ would have been necessary simply to prevent runaway 'icehouse' conditions. Moreover, the glaciation was extremely short-lived⁸, so the available proxies may not pick up a possible reduction in levels of CO₂ during that brief period. CO₂ has a short residence time in the atmosphere, and a transient reduction in atmospheric CO₂ is a viable explanation for the glaciation⁹.

The Jurassic mismatch is a more persistent and problematic feature. There is also a notable mismatch during this time interval between the proxy data, which indicate greatly increased concentrations of CO₂, and more moderate estimates that come from models of the carbon cycle¹⁰. So levels of atmospheric CO₂ may not have been as high as the proxies indicate. Nevertheless, the cool climate of the Jurassic remains something of a mystery.

Clearly, factors other than atmospheric CO₂ can influence climate on geological timescales. Continental drift creates supercontinents and breaks them into smaller fragments, gives birth to mountain belts and high plateaux, and shifts the distribution of continents among latitudinal belts and

hemispheres. All of these factors affect the atmosphere and the ocean circulation, as well as the intensity of the hydrological cycle, which ultimately dictates the cloudiness of the planet and its ice-cap coverage. These changes in the water cycle also alter the reflectivity of Earth as a whole, and can therefore affect global climate.

When we put everything we know into models of the carbon cycle, though, we predict changes in atmospheric CO₂ that largely parallel inferred climate shifts. So the lack of close correspondence between climate change and proxy indicators of atmospheric CO₂ may force us to re-evaluate the proxies, rather than disavow the notion that substantially increased atmospheric CO₂ will indeed lead to marked warming in the future. ■

Human evolution

A start for population genomics

S. Blair Hedges

The most thorough analysis yet of the divergence of sequences in human mitochondrial DNA has been carried out. The results support the view that modern humans originated in Africa.

When and where did our species arise? Over the past two decades molecular evolutionists have vigorously pursued this question. DNA evidence from the cell powerhouse known as the mitochondrion has figured prominently in these studies, with mutations providing the raw data for producing evolutionary trees and molecular clocks to time sequence divergence. The mitochondrial family tree of humans has suggested that our roots lie in Africa^{1–3}, but this tree has had only weak statistical support. Conversely, other researchers have proposed that modern humans arose simultaneously in different regions of the world⁴.

Now, on page 708 of this issue, Gyllenstein and colleagues⁵ describe an analysis of the complete mitochondrial genomes of 53 people of diverse geographical, racial and linguistic backgrounds. At 16,500 base pairs, each sequence is much longer than those previously studied. The upshot is a robust tree rooted in Africa, which times the exodus from Africa to within the past 100,000 years (recent in evolutionary terms). With this result, the pendulum swings further towards the claim that modern humans, *Homo sapiens*, originated in Africa.

Our closest living relatives are African apes, so why is an African origin for modern humans controversial? The reason is that our immediate predecessors in the genus *Homo*, now extinct, are known to have wandered out of Africa as early as two million years ago. The main alternative to an African origin, the multiregional model, holds that modern humans arose simultaneously in Africa,

Lee R. Kump is in the Department of Geosciences and the Astrobiology Research Center, 503 Deike Building, Pennsylvania State University, University Park, Pennsylvania 16802, USA.

e-mail: lkump@psu.edu

1. Walker, J. C. G. & Kasting, J. F. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **97**, 151–189 (1992).
2. Berner, R. A. *Am. J. Sci.* **294**, 56–91 (1994).
3. Frakes, L. A., Francis, J. E. & Syktus, J. I. *Climate Modes of the Phanerozoic* (Cambridge Univ. Press, 1992).
4. Veizer, J., Godderis, Y. & François, L. M. *Nature* **408**, 698–701 (2000).
5. Veizer, J. et al. *Chem. Geol.* **161**, 59–88 (1999).
6. Crowley, T. J. & Baum, S. K. *J. Geophys. Res.* **100**, 1093–1101 (1995).
7. Gibbs, M. T., Barron, E. J. & Kump, L. R. *Geology* **25**, 447–450 (1997).
8. Brechley, P. J. et al. *Geology* **22**, 295–298 (1994).
9. Kump, L. R. et al. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **152**, 173–187 (1999).
10. Berner, R. A. *Science* **276**, 544–547 (1997).

Europe and Asia from these predecessors⁴. Proponents of this view argue that the fossil record indicates transitions between, for example, Neanderthals (*H. neanderthalensis*) and modern humans in Europe, and between *H. erectus* and modern humans in Asia. However, the existence of non-African transitional fossils is debatable^{6,7}, and there is genetic evidence⁸ that Neanderthals did not widely interbreed with modern humans even though the two coexisted for at least 10,000 years. Such coexistence is the strongest evidence for recognizing the two as separate species.

The crux of the mitochondrial evidence for an African origin has been the presence of several African lineages deep in the evolutionary trees, even though they have only had weak statistical support. Gyllenstein's team⁵ also found this pattern, but obtained a robust tree by collecting a larger data set than in previous studies. The three earliest branches in their tree lead exclusively to Africans, and two of the splits are statistically significant. Interpreted literally, the tree indicates that some Africans are closer to Europeans and Asians than to other Africans. However, the history of a single gene or molecule may not always mirror that of the population, and other molecular studies place Africans in a single group⁹. Together, these studies suggest that the founding population leaving Africa carried with it a subset of mitochondrial alleles — alternative forms of the same gene — and that African populations continued to interbreed after the exodus.

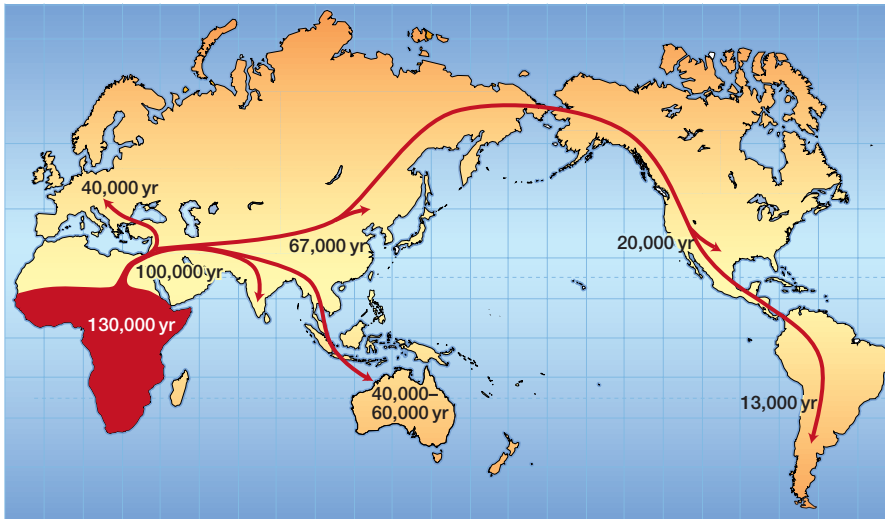


Figure 1 The origin and dispersal of modern humans, *Homo sapiens*. The time of origin of modern humans is not well known but may have been about 200,000 (130,000–465,000) years ago. New evidence from mitochondrial genomes⁵ bolsters the hypothesis that the place of origin was sub-Saharan Africa and that the dispersal from Africa occurred within the past 100,000 years. The earliest known fossil and archaeological evidence on each continent¹⁴, shown on the map, is consistent with this view.

Gyllenstein and colleagues estimate that the divergence of Africans and non-Africans occurred $52,000 \pm 28,000$ years ago, shortly followed by a population expansion in non-Africans. This date may even be a bit too recent. Other genetic markers indicate an exodus from Africa around 100,000 years ago^{9–11}, which would be more consistent with fossil and archaeological evidence of modern humans outside Africa (Fig. 1). But no single genetic marker can time that event precisely, and the mitochondrial date is in the right ballpark. Some nuclear DNA markers have suggested earlier dates for the exodus from Africa¹², so more data are needed to provide a fuller picture. Nonetheless, most of the genetic evidence indicates that there were only about 10,000 breeding individuals for a long time before the recent expansion of modern humans outside Africa¹³. Such a small population size is incompatible with the multiregional model, which would require many more individuals to maintain gene flow among continents.

A different question is when *H. sapiens* arose in the first place. Molecular clocks would be well suited to address that question if our closest relative were living. But our closest relative in the genus *Homo*, whether *H. erectus* or some other species, is unfortunately extinct. The earliest fossils of modern *H. sapiens* are 130,000 years old¹⁴, however, so that is the upper bound on the origin of our species. Studies of ancient DNA provide hints to a lower bound. The split between *H. neanderthalensis* (a species which is not necessarily our closest relative) and *H. sapiens* has been timed by a DNA clock at 465,000 years ago⁸. So our species probably arose somewhere between 130,000 and 465,000 years ago. An estimate of 200,000 years ago

is not unreasonable given the transition seen in the African fossil record between archaic and modern humans around that time¹⁴.

Gyllenstein and colleagues⁵ have used sequences from a large number of complete mitochondrial genomes to address these evolutionary questions, an approach that could be called population genomics. The number of such genome sequences will surely grow rapidly in the near future, and complete

sequences of nuclear genomes, from more than one human, are to be expected. Genes responsible for physical and behavioural traits will probably be found and their allelic histories will provide additional information. Molecular evolutionary trees and time estimates will have greater precision, all of which will help to clarify our evolutionary history.

S. Blair Hedges is in the Department of Biology, Institute of Molecular Evolutionary Genetics, and the Astrobiology Research Center, 208 Mueller Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802, USA.

e-mail: sbh1@psu.edu

1. Cann, R. L., Stoneking, M. & Wilson, A. C. *Nature* **325**, 31–36 (1987).
2. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. *Science* **253**, 1503–1507 (1991).
3. Hedges, S. B., Kumar, S., Tamura, K. & Stoneking, M. *Science* **255**, 737–739 (1992).
4. Wolpoff, M. H. in *The Human Revolution: Behavioral and Biological Perspectives on the Origins of Modern Humans* (eds Mellars, P. & Stringer, C.) 62–108 (Princeton Univ. Press, NJ, 1989).
5. Ingman, M., Kaessmann, H., Pääbo, S. & Gyllenstein, U. *Nature* **408**, 708–713 (2000).
6. Stringer, C. B. & Andrews, P. *Science* **239**, 1263–1268 (1988).
7. Swisher, C. C. I. et al. *Science* **274**, 1870–1874 (1996).
8. Krings, M., Geisert, H., Schmitz, R. W., Krainitzki, H. & Pääbo, S. *Proc. Natl Acad. Sci. USA* **96**, 5581–5585 (2000).
9. Nei, M. & Roychoudhury, A. K. *Mol. Biol. Evol.* **10**, 927–943 (1993).
10. Goldstein, D. B., Linares, A. R., Cavalli-Sforza, L. L. & Feldman, M. W. *Proc. Natl Acad. Sci. USA* **92**, 6723–6727 (1995).
11. Tishkoff, S. A. et al. *Science* **271**, 1380–1387 (1996).
12. Zhao, Z. et al. *Proc. Natl Acad. Sci. USA* **97**, 11354–11358 (2000).
13. Harpending, H. C. et al. *Proc. Natl Acad. Sci. USA* **95**, 1961–1967 (1998).
14. Lewin, R. *Human Evolution* (Blackwell Science, Malden, MA, 1999).

Photonics

On the threshold of success

Richard De La Rue and Chris Smith

Tiny crystals that interact strongly with certain frequencies of light can be used to make semiconductor lasers. They can now be stimulated electrically, and thus be integrated with conventional electronic devices.

Photonic crystals occur naturally as semiprecious opal and the microscopic structures on the wings of some tropical butterflies. They are repeating structures that inhibit the propagation of some frequencies of light. They often produce brilliant colour effects by selecting narrow parts of the spectrum of white light falling on them. If the selection effect is strong enough, the crystal can completely exclude photons of a particular wavelength, creating a three-dimensional 'photonic bandgap'^{1,2}. Any defect in the structure will bottle up the forbidden light — like a tiny optical cavity — by blocking the propagation of light in all directions. To be useful in optoelectronics, the light has to be allowed to escape from this cavity.

Last year, the first semiconductor laser

incorporating a photonic bandgap was created by shining a laser beam on the device to stimulate light emission³. In a paper in *Electronic Letters*, Zhou et al.⁴ now report a photonic bandgap laser that can be excited electrically, allowing it to be incorporated into electronic devices. Such photonic lasers are only a few micrometres in size, opening up new possibilities for optoelectronics and fibre-optical communications.

The photonic-bandgap effect, first proposed just over a decade ago, is the optical equivalent of the 'forbidden energy bandgap' found in the electronic structure of semiconductor crystals. The electronic bandgap excludes electrons of certain energies, an essential feature of semiconductor devices from transistors to lasers. Semiconductor lasers based on electronic bandgaps are well-



100 YEARS AGO

The recent sporting experiences of Mr. Selous. Having seen, or shot, practically every species of great game in South and South-east Africa, the indefatigable author of the volume before us has devoted several seasons in the closing decade of the century to hunting-trips in the northern hemisphere. These expeditions included three trips to Asia Minor in search of the wild goat, the Armenian sheep, and the Asiatic red deer, and two to the Rocky Mountains, where wapiti, mule-deer, white-tailed deer, prongbuck and lynx fell to the practised aim of the veteran hunter. Not that Mr. Selous is by any means merely a hunter; he is likewise an observant field-naturalist and enthusiastic egg-collector, having contributed many years ago an important paper on African antelopes to the *Proceedings of the Zoological Society*, while his recent expeditions to Asia Minor have furnished material for an ornithological paper to the *Ibis*. It is perhaps needless to add that such an experienced hunter may be depended upon not to shoot animals for the mere sake of slaying, and that after obtaining a few fine examples of the species he encountered for the first time to add to his splendid collection at Alpine Lodge, Worplesdon, and occasionally killing an individual or two for the commissariat, Mr. Selous has always been content to stay his hand.

From *Nature* 6 December 1900.

50 YEARS AGO

There is an
MSE centrifuge
model for every
research and routine
requirement in the
modern laboratory
MEASURING & SCIENTIFIC EQUIPMENT, LTD.,
14-28 SPENSER ST., LONDON, S.W.1

From *Nature* 9 December 1950.

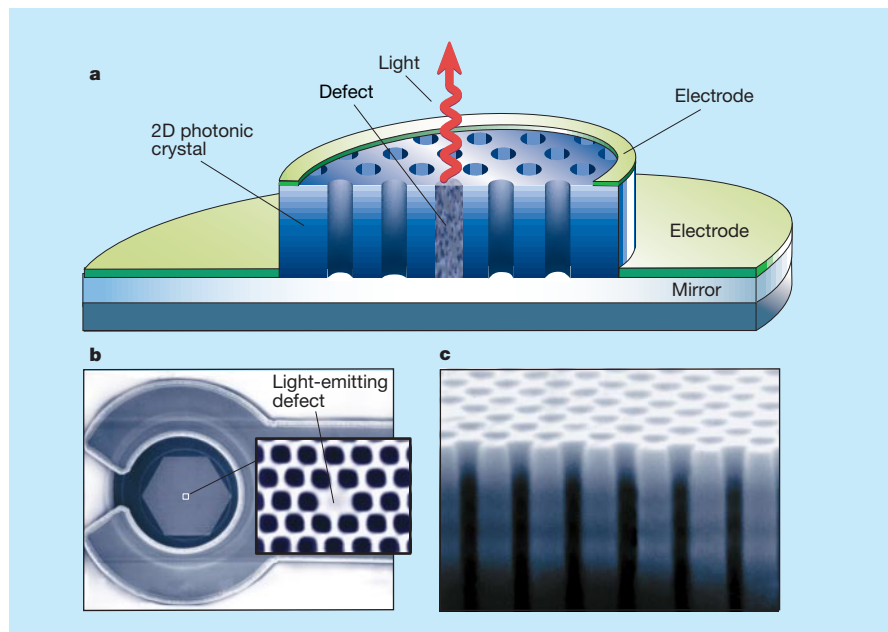


Figure 1 A photonic-bandgap semiconductor laser. **a**, The device built by Zhou *et al.*⁴ emits light following injection of current through the electrodes. The light is emitted from an optical microcavity, created by a single defect in the centre of a two-dimensional photonic crystal. **b**, A scanning electron micrograph of the photonic-bandgap laser from above, showing a single defect approximately 0.4 μm across. **c**, Scanning electron micrograph of a cross-sectional view of the photonic crystal.

established devices, being small, cheap and efficient at converting electric current into light at useful wavelengths. So why go to the bother of making photonic-bandgap lasers that require artificial crystal structures? Ordinary semiconductor lasers are still typically about half a millimetre long, so a photonic-bandgap laser represents a significant reduction in size. This means it can provide an extremely focused source of light, or that many more of them can be packed together in an integrated optical circuit.

In Zhou *et al.*'s device⁴ the ideal (three-dimensional) photonic crystal structure has been modified to become a planar two-dimensional lattice of cylindrical holes, with one small 'defect cavity' — actually a missing hole (Fig. 1). The defect traps photons inside a cavity just 0.4 micrometres wide. The photonic crystal is placed within a mirror, ensuring that photons can escape only from the top of the cavity through the semiconductor–air interface. The efficiency of this laser (measured as power output over current input) is still quite low — a modest 0.0015 W A^{-1} . A larger laser (but still compact at 16 μm long) described by Rennon *et al.*⁵ in the same issue of *Electronic Letters* achieves an efficiency of about 0.04 W A^{-1} , about 25 times better. This means that most of the electrical input in Zhou *et al.*'s device is producing heat rather than light. Both of these values are significantly inferior to those for existing semiconductor lasers, where efficiencies of 0.4 W A^{-1} are standard.

The threshold current — the level at which the laser switches on — of the photonic-

bandgap laser designed by Zhou *et al.*⁴ is 300 μA , much more than the threshold of 40 μA and lower seen in other semiconductor lasers^{6,7}. There is still some way to go before technology delivers the full potential of the photonic-bandgap semiconductor laser. But we can expect such lasers to switch on at currents as small as 1 μA and produce a maximum of about 10 μW of light, together with a power-conversion efficiency of 50% or higher. Although such microwatt output power levels seem small compared with the several milliwatts obtained with ordinary semiconductor lasers, the corresponding number of photons per second is still large enough to give adequate performance for practical fibre-optical systems. This type of laser could eventually have a near-zero threshold.

So what are the prospects for photonic-bandgap lasers? If they reach their full potential, arrays of these devices might be used in high-speed optical connections between adjacent silicon chips in a computer — or even between different points on the same chip. In these applications, heat production is likely to be a severe problem, because the devices will be packed closely together. So small photonic-bandgap lasers (possibly just a square micrometre in size) that emit light in a range of directions and switch on at very low current levels will be an attractive choice. The power needed per device reduces with size, but efficient conversion between electrons and photons will remain an essential feature of these tiny lasers if they are to operate at room temperatures

without the need for bulky cooling arrangements⁸. Another possible application for arrays of photonic lasers would be in multi-colour light-emitting displays, in which the colour (wavelength) of the light 'pixels' could be varied easily.

The possible applications of photonic-bandgap structures are not restricted to telecommunications. The ability to control the emission process could lead to lasers that have intrinsically quantum-optical properties — for example, single-photon sources that could be used in experiments to explore the quantum nature of light. Such quantum lasers could find eventual applications in quantum cryptography and optical quantum computers. And if a laser with a near-zero threshold could be produced, then the 'spare' heat from the human body

might provide enough energy to switch it on, suggesting considerable economic and ecological implications. The work reported by Zhou *et al.*⁴ and others is an important step in this direction — and the long-term prospects for photonic-bandgap lasers are surely bright.

Richard De La Rue and Chris Smith are with the Optoelectronics Research Group, Department of Electronics and Electrical Engineering, University of Glasgow, Glasgow G12 8QQ, UK.
e-mail: r.delarue@elec.gla.ac.uk

1. Yablonovitch, E. *Phys. Rev. Lett.* **58**, 2059–2062 (1987).
2. John, S. *Phys. Rev. Lett.* **58**, 2486–2489 (1987).
3. Painter, O. *et al. Science* **284**, 1819–1821 (1999).
4. Zhou, W. D. *et al. Electron. Lett.* **36**, 1541–1542 (2000).
5. Rennon, S. *et al. Electron. Lett.* **36**, 1548–1550 (2000).
6. Baba, T. *et al. Electron. Lett.* **36**, 790–791 (2000).
7. Huffaker, D. L. *et al. IEEE Photon. Technol. Lett.* **8**, 974–976 (1996).
8. De La Rue, R. M. *Science* **286**, 2459–2460 (1999).

and stomatal index². (The stomatal index is given by $[\text{stomatal density}/(\text{stomatal density} + \text{epidermal-cell density})] \times 100$, where epidermal cells include guard cells as well as others that constitute the outer covering of leaves.) But several experiments have shown that enrichment of the present atmospheric CO₂ levels induces only minor further changes in stomatal density^{2,3} and index². The implication is that the response of the stomatal index to CO₂ in many plant species is now close to saturation. It seems that plants may continue to make enough stomata, regardless of how much more CO₂ is dumped into the atmosphere by humans.

The results of Gray *et al.*¹ lend support to this view. The authors' laboratory experiments show that roughly doubling the current atmospheric CO₂ levels has no effect on the stomatal index in the leaves of at least one strain of *Arabidopsis*. They also show that plants with a loss-of-function mutation in the *HIC* gene exhibit a large increase in this index when grown at the higher CO₂ concentration. The implication is that the unmutated *HIC* gene represses this increase in stomatal index in response to atmospheric CO₂ enrichment. An increase in stomatal density might well be as much of a problem as a decrease, because water loss might be too high.

So, what does *HIC* do? Mutant plants with defects in stomatal patterning have been identified previously^{4–7}. But, unlike

Plant biology

Coping with human CO₂ emissions

Laura Serna and Carmen Fenoll

For two centuries, a natural experiment has been showing how increasing atmospheric CO₂ affects plants. Laboratory work provides pointers as to how they will respond in the future.

The industrial revolution in the western world had enormous socioeconomic effects — but it also unwittingly started a global experiment that has now been running for two centuries. One of the unforeseen consequences of the industrial revolution was a steady increase in the amount of carbon dioxide in the atmosphere. Plants responded by having fewer stomata — paired 'guard' cells, found mainly on the surface of leaves, that control the uptake and release of gases. Enrichment of atmospheric CO₂ continues. But how will plants respond in the future? Water loss and CO₂ uptake occur primarily through the stomata, so a drastic further decrease in their numbers would impair processes such as transpiration and photosynthesis.

On page 713 of this issue¹, Gray *et al.* tackle this problem. They show that the small mustard plant *Arabidopsis thaliana* is endowed with a gene, christened *HIGH CARBON DIOXIDE* (*HIC*), that prevents changes in the number of stomata in response to further atmospheric CO₂ enrichment. This finding not only provides new evidence linking CO₂ and stomatal development, but also opens the way to understanding how plants will cope with anthropogenic CO₂ emissions in the future.

An inverse relationship between stomatal density and atmospheric CO₂ levels, from pre-industrial to modern times, has been demonstrated by analyses of specimens collected from herbaria over the past 200 years². Laboratory experiments in which plants

were grown at pre-industrial CO₂ levels confirmed the relationship between CO₂ concentration and both stomatal density^{2,3}

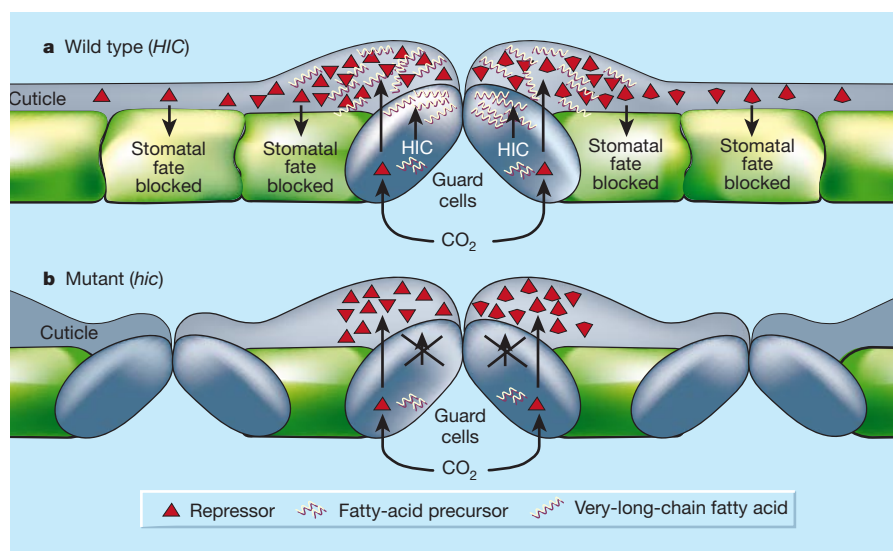


Figure 1 Plant stomata and anthropogenic carbon dioxide emissions. Over the past 200 years, the number of plant stomata (each of which consists of a pair of guard cells, delimiting a pore for gas exchange) has decreased in response to increasing CO₂ levels^{2,3}. This trend may not continue in the future. Gray *et al.*¹ show that the stomatal index remains the same in normal *Arabidopsis* plants in response to CO₂ concentrations double those of today, but increases in plants with a mutant *HIC* gene. The authors suggest that the normal *HIC* gene negatively regulates stomatal development at the higher CO₂ concentrations, as shown here. a, The normal *HIC* gene may encode an enzyme involved in the synthesis of very-long-chain fatty acids, deposited at the surface of the guard cells. These fatty acids might be needed for the diffusion of a molecule — stimulated by high CO₂ concentrations — that can reach neighbouring epidermal cells and repress their ability to develop into stomata. b, In *hic* mutant plants, the relevant fatty acids are not made. Diffusion of the repressor is therefore impaired, so it cannot reach all its target cells, which develop into stomata.

these plants, the *hic* mutants do not have stomata that directly contact one another¹. So it seems that the normal *HIC* gene is not required to ensure that a full complement of non-stomatal cells surrounds each stoma.

Gray *et al.* next cloned the *HIC* gene, and found that its sequence has many bases in common with the *Arabidopsis KCS1* gene, which codes for a 3-keto acyl coenzyme A synthase⁸. In plants, such enzymes are involved in synthesizing fatty acids with very long chains, which are important in the cuticle waxes on the surface of epidermal cells⁹. The authors worked out the amino-acid sequence of the normal and mutant *HIC* proteins, and propose that the mutant protein should not be able to synthesize a component or components of the waxy cuticle of guard cells. They suggest that these alterations might result in an increase in the stomatal index in response to atmospheric CO₂ enrichment.

But how might an altered cuticle influence epidermal development in such a way that the stomatal index is increased? Gray *et al.* put forward a possible model (Fig. 1): in the mutant plants, the change in the cuticle overlying the guard cells interferes with its permeability to a hypothetical substance that is stimulated by doubled CO₂ levels, blocking the effects of this substance. In normal plants this substance might diffuse from the cuticle above the guard cells to nearby epidermal cells, there repressing the development of extra stomata. Very limited diffusion of the repressor might occur in *hic* mutant plants, and this might be sufficient to prevent the formation of stomata adjacent to one another. As the authors point out¹, if this model fits reality, we might be close to showing that a lateral inhibition process operates during the establishment of stomatal patterns at CO₂ levels that are double those of today.

To investigate further whether cuticular waxes influence stomatal development, Gray *et al.* measured the stomatal index of *eceriferum-1* and *eceriferum-6* mutants, which have alterations in wax composition¹⁰. Both mutants have higher stomatal indices than do wild-type plants at current CO₂ atmospheric levels¹. This finding establishes a link between the lipid composition of the cuticle and the development of stomata. Work on these and other mutant plants is already revealing that cuticle lipids are important in processes involving cell–cell interactions at plant surfaces¹¹. But the question of whether these lipids simply help these processes — providing an adequate environment for other molecules to operate — is still open.

The main message of Gray *et al.*'s work¹ is that plants seem to be well armed to cope with a further enrichment in atmospheric CO₂ (at least until CO₂ levels are double those at present). Genes such as *HIC* should

ensure that, at high CO₂ concentrations, changes in stomatal indices are kept to a minimum. But *HIC* does not seem to be necessary to regulate stomatal development at current CO₂ levels, as mutant *hic* plants showed no stomatal defects at such concentrations¹. Future challenges include not only unveiling the components of the *HIC*-mediated developmental signalling pathway, but also understanding the mechanisms that allowed plants to decrease their number of stomata as CO₂ levels rose from pre-industrial to present-day levels. In any case, it seems that further increases in CO₂ emissions might not drastically alter stomatal density. But be on guard — the global experiment is still running. ■

Geochemistry

Mantle recycled in Sardinia

Barry B. Hanan

The origin of the 'EM-1 signature' evident in certain basalt rocks is a long-standing puzzle. Comparison of data from Pitcairn Island and Sardinia may shed some light on the enigma.

Oceanic 'large igneous provinces' (LIPs) are areas of the sea floor where the ocean crust is especially thick — 30 km or more. During the plate tectonic process of subduction along island arcs, when sea floor sinks back down into the mantle, LIPs are thought to remain at the surface because of their comparative buoyancy and to accrete to the continental crust. On page 701 of this issue, however, Gasperini *et al.*¹ propose that at least some LIP ocean floor has been recycled back into the deep mantle during plate subduction and has reappeared in the Logudoro lavas of northwestern Sardinia. This work bears upon a long-standing geochemical puzzle: the origin of the 'EM-1 signature' in certain rocks.

Large igneous provinces (such as the Ontong–Java plateau) and 'hotspot' volcanoes (such as Pitcairn, Tristan da Cunha, Iceland and Hawaii) both form from plumes of hot material rising up through Earth's mantle. Mantle plumes consist of a huge head (500–2,000 km in diameter) on top of a tail (100 km wide) of hot upwelling deep mantle. Ocean LIPs are thought to result from lavas erupted during melting of the head in an early stage of plume development, producing massive amounts of lava the size of continents. By contrast, hotspot volcanoes form over more mature plumes, after dispersion of the plume head, by melting of plume tails.

Normal ocean crust is created at mid-ocean spreading centres, where upwelling mantle decompresses and melts. This crust, 4–8 km thick, consists of pillow basalts — lavas chilled by sea water — that are fed from below by dykes emanating from a magma

Laura Serna and Carmen Fenoll are in the Facultad de Ciencias del Medio Ambiente, Universidad de Castilla-La Mancha, E-45071 Toledo, Spain.

e-mails: lserna@amb-to.uclm.es

cferoll@amb-to.uclm.es

1. Gray, J. E. *et al.* *Nature* **408**, 713–716 (2000).
2. Woodward, F. I. *Nature* **327**, 617–618 (1987).
3. Woodward, F. I. & Kelly, C. K. *New Phytol.* **131**, 311–327 (1995).
4. Schneider, K., Wells, B., Dolan, L. & Roberts, K. *Development* **124**, 1789–1798 (1997).
5. Yang, M. & Sack, F. D. *Plant Cell* **7**, 2227–2239 (1995).
6. Wei, N. *et al.* *Plant Cell* **6**, 629–643 (1994).
7. Berger, D. & Altmann, T. *Genes Dev.* **14**, 1119–1131 (2000).
8. Todd, J., Post-Beittenmiller, D. & Jaworski, J. G. *Plant J.* **17**, 119–130 (1999).
9. Post-Beittenmiller, D. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 405–430 (1996).
10. Jenks, M. A. *et al.* *Plant Physiol.* **108**, 369–377 (1995).
11. Mariani, C. & Wolters-Arts, M. *Plant Cell* **12**, 1795–1798 (2000).

chamber and are overlain by deep-sea sediment. Oceanic LIPs, which may constitute up to 10% of the ocean floor, have a similar structure to normal ocean crust but are as much as five times thicker².

The Earth's mantle contains several isotopically 'extreme' components that contribute to the sources of ocean basalts^{3,4}. Three of these components have been identified in the radiogenic isotope signatures of hotspot volcanoes. They are known as HIMU (high ²³⁸U to ²⁰⁴Pb) and EM-1 and EM-2 — EM standing for 'enriched mantle', meaning enriched in certain elements. These components are thought to represent ancient recycled oceanic crust and its overlying sediment that has undergone varying degrees of hydrothermal alteration on the sea floor (see ref. 5 for review).

Gasperini *et al.*¹ point out that the geochemistry of the Logudoro lavas resembles that of Pitcairn, an EM-1-type hotspot island in the Pacific Ocean. The EM-1 mantle source is especially puzzling because its geochemistry is consistent not only with contributions from recycled, ancient crust and deep-sea sediment⁶, but also with ancient lithosphere from beneath the continents⁷. Because EM-1 signatures are evident in LIP basalts associated with plume heads rising beneath the continents⁷, there is the question of whether the EM-1 signature is due to contamination of plume heads by subcontinental lithosphere or whether it derives from the original plume source. Solving this enigma is fundamental to deciphering the origin of the isotope components of the mantle⁸.

The Pitcairn and Logudoro lavas both

have Hf–Nd isotope compositions that indicate that their mantle source cannot have contained much subducted deep-sea sediment. The Logudoro lavas have higher amounts of Ba, Sr, Eu and Pb, with trace-element patterns similar to those seen in gabbro rocks enriched in the mineral plagioclase. The elements Ba, Sr, Eu and Pb preferentially partition into plagioclase during the magmatic processes in the shallow mantle that form gabbro in the lower part of the oceanic crust. The major element compositions of the Logudoro lavas rule out contamination of the melts as they travelled through the present-day Sardinian crust, so a mantle source for the plagioclase trace-element signature is required. Plagioclase is not stable at high pressure beneath the thick continental crust (35–45 km), so it is difficult to explain the EM-1 signature as a consequence of subcontinental lithosphere being swept up by rising plume material.

Likewise, the very low $^{187}\text{Os}/^{188}\text{Os}$ ratios in the subcontinental mantle inclusions known as xenoliths⁹ suggest that EM-1 signatures are unlikely to have originated in the subcontinental lithosphere. Basalts have a much higher Re/Os ratio than their mantle source. This results in a rapid radiogenic increase in the $^{187}\text{Os}/^{188}\text{Os}$ ratio in ocean crust and provides a robust tracer for recycled basalt in plume sources. HIMU, EM-1 (including Pitcairn) and EM-2 hotspot basalts all have high $^{187}\text{Os}/^{188}\text{Os}$ ratios, pointing to the existence of large proportions of recycled basalt in their mantle sources^{10,11}.

The Logudoro and Pitcairn lavas have the EM-1 signature in common; they also have similar Pb, Sr, Nd and Hf isotopic compositions, and high Th/U ratios relative to the normal basalts formed at mid-ocean ridges. Given these characteristics, and the shallow mantle plagioclase trace-element signature of the Logudoro lavas, Gasperini *et al.*¹ argue that the lavas are derived from deep mantle-plume sources of LIP origin. They account for the differences in the abundances of Ba, Sr and Pb by arguing that there is more recycled LIP lower crust than upper crust in the Logudoro relative to the Pitcairn lavas.

Finally, there is another possible explanation for the occurrence of the EM-1 signature. The signature might stem from the lower Mg-rich silicate part of the continental crust, which has the appropriate unradiogenic Pb, high Th/U and radiogenic Os composition^{12,13}. Although the continental crust is 20–30 km thick beneath northwestern Sardinia, Gasperini *et al.* rule out this possibility for the Logudoro lavas. This is because basalt of the same age in the Tyrrhenian Sea, with no intervening continental crust, has the same geochemical signature as Logudoro.

The combination of radiogenic isotope ratios and trace-element patterns provides a

powerful tool for tracing the mantle sources of basalts and accounting for their variety. Geochemists have made good progress in matching such mantle components to geodynamic phenomena. Continued application of these tools should eventually provide a geochemical database to solve the puzzle of the EM-1 signature. ■

Barry B. Hanan is in the Department of Geological Sciences, San Diego State University, 5500 Campanile Drive, San Diego, California 92182-1020, USA.

e-mail: barry.hanan@sdsu.edu

1. Gasperini, D. *et al.* *Nature* **408**, 701–704 (2000).
2. Neal, C. R., Mahoney, J. J., Kroenke, L. W., Duncan, R. A. &

- Petterson, M. G. in *Large Igneous Provinces: Continental, Oceanic, and Planetary Flood Volcanism* (eds Mahoney, J. J. & Coffin, M. F.) 183–216 (American Geophysical Union, Washington DC, 1997).
3. White, W. M. *Geology* **13**, 115–118 (1985).
4. Zindler, A. & Hart, S. R. *Annu. Rev. Earth Planet. Sci.* **14**, 493–571 (1986).
5. Hofmann, A. W. *Nature* **385**, 219–229 (1997).
6. Weaver, B. L. *Earth Planet. Sci. Lett.* **104**, 381–397 (1991).
7. Mahoney, J. J., Nicollet, C. & Dupuy, C. *Earth Planet. Sci. Lett.* **104**, 350–363 (1991).
8. Hanan, B. B. & Graham, D. W. *Science* **272**, 991–995 (1996).
9. Pearson, D. G. *et al.* *Nature* **374**, 711–713 (1995).
10. Hauri, E. H. & Hart, S. T. *Earth Planet. Sci. Lett.* **114**, 353–371 (1993).
11. Reisberg, L. *et al.* *Earth Planet. Sci. Lett.* **120**, 149–167 (1993).
12. Chesley, J. T. & Ruiz, J. *Earth Planet. Sci. Lett.* **154**, 1–11 (1998).
13. Rudnick, R. L. & Goldstein, S. L. *Earth Planet. Sci. Lett.* **98**, 192–207 (1990).

Immunology

A Toll for DNA vaccines

Robert L. Modlin

The mammalian protein that triggers an innate immune response to bacterial DNA has been identified. This discovery may help in understanding and improving DNA-based vaccines.

Innate immunity enables an organism to respond rapidly to invading microorganisms. To do this, the innate immune system uses receptor proteins that can recognize a microbial pathogen by the molecular patterns it displays. There are many pattern-recognition molecules that are involved in innate responses to microbial lipids and carbohydrates. Bacterial DNA is also a potent inducer of innate immunity, but how is such DNA recognized and distinguished from the DNA of the host organism? On page 740 of this issue¹, Hemmi and colleagues provide an answer. They have identified a mammalian protein — Toll-like receptor 9 (TLR9) — that recognizes a specific pattern in bacterial DNA.

The Toll family of pattern-recognition receptors has been conserved over hundreds of millions of years of evolution, occurring in species from insects to humans^{2,3}. We already know much about innate immunity from studies of Toll proteins in the fruitfly *Drosophila melanogaster*⁴. First, *Drosophila* Toll proteins show exquisite specificity for the molecular pattern, or ligand, to which they respond, with Toll itself being responsible for mediating responses to fungi and Gram-positive bacteria, and the so-called 18-wheeler protein mediating activation by Gram-negative bacteria. Second, different proteins of the *Drosophila* Toll family activate distinct intracellular signalling pathways. Third, the different *Drosophila* Toll receptors lead to the production and activation of specific antimicrobial proteins to kill pathogens. *Drosophila* Toll is a savvy exterminator.

It is becoming clear that, as in *Drosophila*, the mammalian Toll-like receptors also display remarkable specificity for different ligands. TLR4, for instance, is required for

the response to lipopolysaccharide⁵, found in the outer membranes of Gram-negative bacteria; TLR2 mediates the response to microbial lipoproteins^{6,7}; and TLR6 cooperates with TLR2 in detecting a subset of bacterial peptidoglycans⁸ (Fig. 1, overleaf).

Hemmi *et al.*¹ add a further piece to the puzzle. They show that a specific pattern in bacterial DNA — consisting of an unmethylated pair of nucleotides, cytosine and guanosine (CpG), plus their flanking regions — activates mouse immune cells via TLR9. Their striking finding is that when the TLR9 gene in mice is selectively inactivated, the animals' immune cells do not proliferate or release protective cytokine proteins in response to CpG-containing DNA. Unmethylated CpG sequences are 20 times more common in bacterial than in mammalian DNA⁹, so mammalian TLR9 is much more likely to be activated by bacterial than by mammalian DNA. Future studies should clarify how TLR9 recognizes this pattern, find out whether other nucleotide patterns also activate this protein, and identify the precise ligands for the other mammalian Toll-like receptors.

Are there any clues to how bacterial DNA activates TLR9? For example, does the DNA bind directly to TLR9, or does it trigger an intermediate step, with TLR9 ultimately being activated by a cellular ligand? The same question also applies to other microbial molecules. Studies in *Drosophila* indicate that pathogens may drive serine proteases — protein-cleaving enzymes — to release Spaetzle, a cellular ligand that activates *Drosophila* Toll¹⁰. But no mammalian equivalent of Spaetzle has yet been found. Preliminary studies indicate, however, that microbial

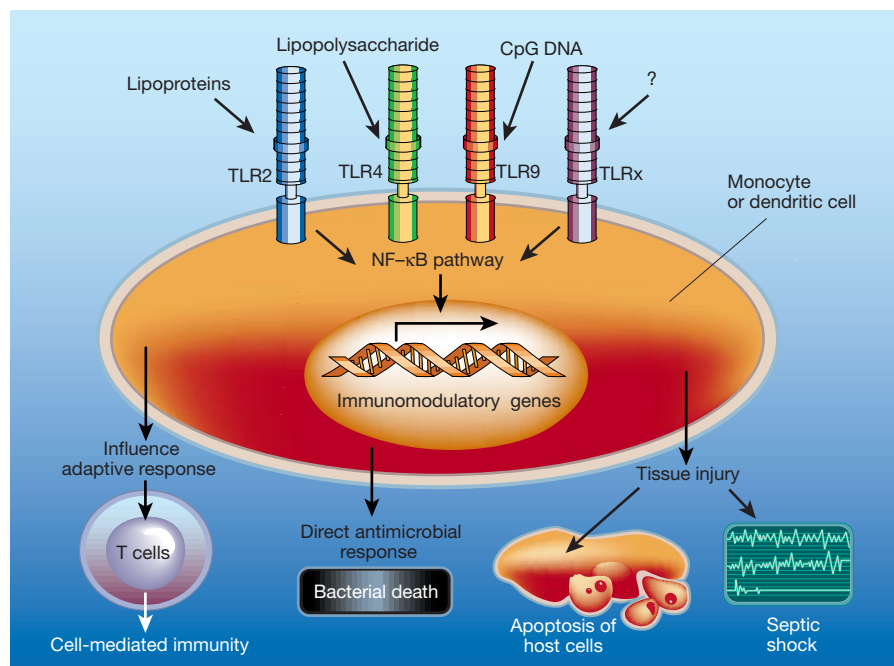


Figure 1 Role of Toll-like receptors in the mammalian innate immune response to pathogens. Mammalian Toll-like receptors are expressed on a variety of immune cells, including monocytes and dendritic cells. Microbial lipoproteins activate mammalian immune cells through Toll-like receptor 2 (TLR2). Lipopolysaccharide activates these cells via TLR4. As shown by Hemmi *et al.*¹, bacterial DNA sequences containing unmethylated cytosine–guanosine dinucleotides (CpGs) work through TLR9. The specificity of the other six mammalian Toll-like receptors is under investigation. Activation of Toll-like receptors kicks off signalling pathways that activate the transcription factor NF-κB, resulting in transcription of genes that modulate and mediate immune responses. One result of these pathways is the release of pro-inflammatory cytokines, which have a say in the adaptive T-cell immune response. Another outcome (at least in fruitflies) is the activation of antimicrobial pathways that directly kill the pathogen. But the activation of Toll-like receptors can also be detrimental to the host. It can contribute to tissue injury in the form of apoptosis (programmed cell death) and the life-threatening symptoms of septic shock. During septic shock, infection leads to a failure of the circulatory system to supply sufficient nutrients and oxygen to tissues, and to remove metabolic wastes from those tissues.

ligands associate directly with mammalian Toll-like receptors¹¹.

If microbial ligands and Toll-like receptors were to interact directly, one would expect this to occur on the cell surface, where Toll-like receptors are found. But, to be active, CpG-containing DNA sequences must be taken up into immune cells by endocytosis^{12,13}. This is a process by which a cell's outer membrane scoops up extracellular substances and transports them into the cell in membrane-bounded vesicles, which fuse together. Toll-like receptors are also taken into cells within vesicles during pathogen uptake¹⁴. So the uptake of CpG DNA sequences by endocytosis might be a way to transport these sequences to an intracellular vesicle containing TLR9. This might also be a clever mechanism by which Toll-like receptors can detect ligands from pathogens residing in vesicles inside cells. But it is not yet known whether the intracellular vesicle is the optimal site for interactions between ligands and Toll-like receptors. Another possibility is that membrane attached signalling factors are recruited to such vesicles to start the activation process.

The greatest practical importance of

Hemmi *et al.*'s paper will probably lie in understanding and improving DNA-based vaccines. The effectiveness of such vaccines depends on two features¹⁵. First, the DNA must encode and express an antigen, such as a bacterial protein, that stimulates the proliferation of responsive T cells, part of the adaptive immune system. Second, CpG sequences in the DNA act as an 'adjuvant'. This means that, by activating the innate immune system to instruct and amplify the adaptive response, these sequences enhance the effectiveness of DNA vaccines.

Hemmi *et al.* show that TLR9 is required for CpG sequences to induce monocytes and dendritic cells to produce interleukin-12. This cytokine is involved in activating T-helper-1 cells, which are needed for protection against many intracellular pathogens. This is, therefore, a mechanism for the adjuvant activity of DNA vaccines. So vaccination techniques, such as gene guns, that insert the DNA vaccine directly into target cells may not be ideal — such techniques might bypass TLR9, which could result in a bias towards a T-helper-2 response (which helps to ward off parasites such as flatworms

Daedalus

David Jones

David Jones, author of the Daedalus column, is indisposed.

and roundworms) rather than the CpG-induced T-helper-1 response. In fact, this has been observed in mouse models.

DNA vaccines tend to be more effective in mice than in humans. But the bacterial DNA sequences that activate mouse and human immune cells differ by only a single nucleotide — the optimal sequence for activating mouse cells is GACGTT; for human cells, it is GTCGTT^{9,16}. The discovery of TLR9 may explain the difference in effectiveness: only 75.5% of the amino acids in the mouse protein are found in the human receptor¹. It will be intriguing to find out how the amino-acid differences contribute to the specificity of recognition of CpG sequences. It may also become possible to create CpG-containing sequences that are more active in humans, acting as better adjuvants to DNA vaccines.

The activation of mammalian Toll-like receptors clearly triggers beneficial immune responses. Toll proteins in the fly also induce signalling pathways that kill microbes directly, although it is not known whether their mammalian counterparts do the same. But the activation of Toll-like receptors might also lead to cell and tissue damage — indeed, it can induce the death of activated cells and contribute to the development of the potentially fatal complication of infections known as septic shock (Fig. 1). Before CpGs and other molecules that activate Toll-like receptors can be truly useful clinically, it will be important to find ways to prevent these harmful side effects. We should then be able to arm ourselves against the ever-increasing threats from infectious predators.

Robert L. Modlin is in the Division of Dermatology, Department of Microbiology and Immunology, and Molecular Biology Institute, UCLA School of Medicine, Los Angeles, California 90095, USA. e-mail: rmodlin@mednet.ucla.edu

- Hemmi, H. *et al.* *Nature* **408**, 740–745 (2000).
- Medzhitov, R., Preston-Hurlburt, P. & Janeway, C. A. Jr *Nature* **388**, 394–397 (1997).
- Hoffmann, J. A. *et al.* *Science* **284**, 1313–1318 (1999).
- Lemaitre, B., Reichhart, J. M. & Hoffmann, J. A. *Proc. Natl. Acad. Sci. USA* **94**, 14614–14619 (1997).
- Poltorak, A. *et al.* *Science* **282**, 2085–2088 (1998).
- Brightbill, H. D. *et al.* *Science* **285**, 732–736 (1999).
- Aliprantis, A. O. *et al.* *Science* **285**, 736–739 (1999).
- Ozinsky, A. *et al.* *Proc. Natl. Acad. Sci. USA* **97**, 13766–13771 (2000).
- Krieg, A. M. *Curr. Opin. Immunol.* **12**, 35–43 (2000).
- Levashina, E. A. *et al.* *Science* **285**, 1917–1919 (1999).
- Yang, R. B. *et al.* *Nature* **395**, 284–288 (1998).
- Hacker, H. *et al.* *EMBO J.* **17**, 6230–6240 (1998).
- Yi, A.-K. *et al.* *J. Immunol.* **160**, 4755–4761 (1998).
- Underhill, D. M. *et al.* *Nature* **401**, 811–815 (1999).
- Gurunathan, S., Klinman, D. M. & Seder, R. A. *Annu. Rev. Immunol.* **18**, 927–974 (2000).
- Hartmann, G., Weiner, G. J. & Krieg, A. M. *Proc. Natl. Acad. Sci. USA* **96**, 9305–9310 (1999).

Quantum theory's last challenge

Giovanni Amelino-Camelia

Quantum theory is 100 years old and still going strong. Combining general relativity with quantum mechanics is the last hurdle to be overcome in the 'quantum revolution'.

This year we celebrate 100 years of quantum theory, and in particular the anniversary of an announcement made by Max Planck at a meeting of the German Physical Society on 14 December 1900. Planck was interested in the nature of radiation emitted by hot objects, and in 1900 he devised a theory that described all of the experimental evidence, but that required a radical new concept: energy is not emitted or absorbed continuously, but in discrete amounts, called quanta. At the time, Planck was not aware of the profound consequences of his work, but gradually physicists realized that they needed quantum concepts to understand the structure of all matter and radiation.

Over the past century there have been many successful tests of quantum mechanics (see accompanying commentary by Anton Zeilinger¹ on page 639 of this issue). Experiments have confirmed even some of the most counterintuitive predictions of quantum theory, including particle-wave duality — the idea that a particle should be treated as a wave. But some physicists still wonder whether quantum theory is a truly fundamental ingredient of the laws of nature, or just a convenient description of some aspects of the microscopic world. It is still possible that quantum mechanics is an approximation to a more fundamental theory, just as newtonian gravity is a special case of the more accurate description of gravity and the relationship between space and time provided by Einstein's theory of general relativity.

The biggest challenge to accepting quantum mechanics as a fundamental theory of nature is that — despite 70 years of attempts² — it has still not been integrated with the classical theory of general relativity. Most of the time, the subtleties of quantum mechanics can be safely ignored by general relativity: gravity drives the expansion of the Universe and the formation of galaxies, whereas quantum theory reigns supreme at the atomic scale. But there are times when quantum mechanics cannot be excluded. For example, a theory unifying gravity and quantum mechanics is required to understand the 'Big Bang' — the first few moments of the Universe when gravitational interactions were very strong and the scales involved were all microscopic.

Another area of concern is the lack of experimental evidence on the interplay between quantum theory and general rela-

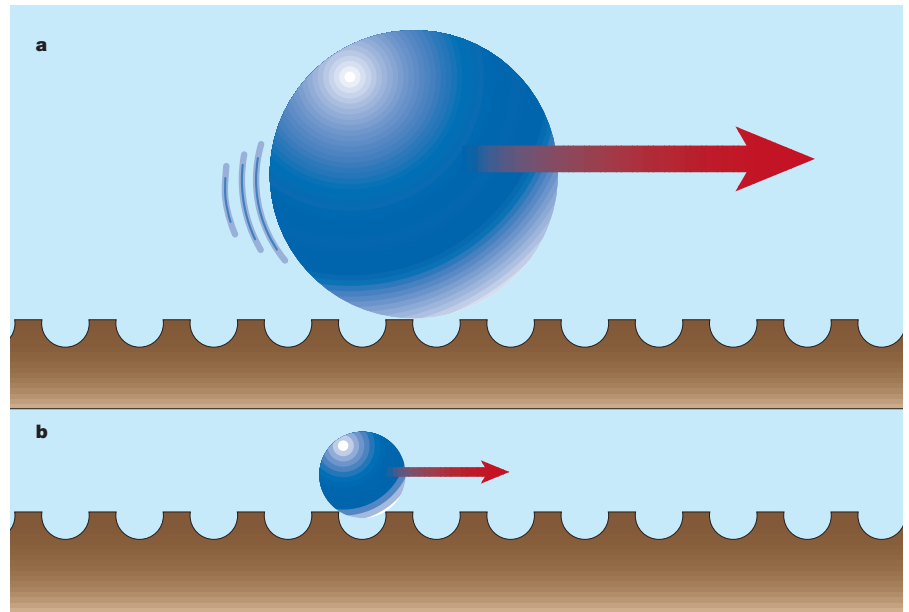


Figure 1 The description of space–time put forward by Einstein's theory of general relativity might eventually reveal quantum behaviour if we observe its effects at extremely small scales. For example, the small-scale structure of a quantum space–time could affect the propagation of particles in a way that depends on their size. This is similar to the effect the microscopic roughness of the surface of a table has on the roll of a marble. In a, the path of the large ball is not perturbed by the roughness of the table, whereas in b, the movement of the smaller ball is severely affected by the nature of the underlying surface.

tivity. The structure of the two theories does allow for situations in which neither can be neglected, but it is extremely hard to create the required conditions in a laboratory. This is crucial if we are to test any of the theoretical ideas that propose to unify quantum mechanics and general relativity.

The incomplete revolution

The 'relativity revolution' and the 'quantum revolution' are among the greatest successes of twentieth-century physics, yet the theories they produced appear to be fundamentally incompatible. General relativity remains a purely classical theory: it describes the geometry of space and time as smooth and continuous, whereas quantum mechanics divides everything into discrete chunks. The predictions of the two theories have been confirmed in a large number of experiments, but each of these experiments is relevant for only one or the other of the two theories because of the different scales involved. Physicists' discomfort with this situation is perhaps best described by Carlo Rovelli's characterization³ of the 1900s as "the century

of the incomplete revolution", suggesting there may be a greater revolution to come.

Part of the underlying theoretical incompatibility arises from the way the two theories treat the geometry of space and time. In quantum mechanics, space–time has the role of a fixed arena within which one describes the evolution of various 'quantum observables', such as the positions of particles. But in general relativity, space and time are dynamical quantities that can respond to and influence other physical processes, such as the movements of planets. So, in principle, the particles in a quantum-mechanics experiment affect the evolution of space–time through the gravitational fields generated by their energies. In practice, the particles are usually so small that they have a negligible effect on space–time dynamics, and general relativity can be ignored. This has been the case in all successful tests of quantum mechanics.

Approaches to quantum gravity

The search for a quantum theory compatible with general relativity can take two different paths. The most drastic option is to look for

Box 1 The Planck length

The modern description of radiation emitted by hot objects is based on the idea of 'quantized' energy: any physical system can access only a discrete series of energy levels. The gap between the allowed energy levels is governed by a fundamental constant, the 'Planck constant' h , which has the experimental value

$h = 6.626 \times 10^{-34}$ J s. By combining h with the speed of light c (2.998×10^8 m s $^{-1}$) and the gravitational constant G (6.673×10^{-11} N m 2 kg $^{-2}$), one gets a length scale $L_p \sim \sqrt{(hG/2\pi c^3)} \sim 10^{-33}$ cm, now called the 'Planck length'. Because L_p is proportional to both G , which sets the strength of gravitational interactions, and

h , which sets the strength of quantum effects, we expect L_p to indicate the strength of effects involving both general relativity and quantum mechanics. Current experiments operate at much larger scales than the minute Planck length, making the detection of any effects that might emerge at this length scale extremely difficult. **G.A.-C.**

an alternative theory based on principles that are profoundly different from everything we know now. In such a theory, quantum mechanics becomes meaningful only as an approximation^{4,5}. The other solution is to maintain the existing principles of quantum mechanics but to develop a more fundamental theory that can answer questions we cannot answer now³ — for example, what happens when space–time is no longer fixed. The situation in which quantum mechanics is abandoned at the level of the more fundamental theory is a much harder challenge, because one would need to start from scratch and no serious candidate has yet emerged. The second option is more easily explored: at least one has quantum theory as a starting point.

There are currently two mature theories based on quantum mechanics that attempt to unify general relativity and quantum mechanics — 'canonical quantum gravity'^{3,6} and 'superstring theory'^{7,8}. Although they share a common goal, these two theories are quite different in the way they approach the technical and physical problems that emerge when building a quantum picture of gravity. In particular, they differ in the way they handle the mathematical infinities that naturally occur in quantum descriptions of gravitational fields. These infinities are not peculiar to quantum gravity: they also appear in the quantum description of electromagnetic fields (the subject of quantum electrodynamics or QED), but here they can be removed by a technique known as perturbative renormalization. Thanks to this technique, QED has become one of the most accurate and best-tested theories of modern physics.

But perturbative renormalization does not work for gravitational fields: the mathematical infinities are too persistent. In canonical quantum gravity, new mathematical techniques (non-perturbative renormalization) have been tried, with mixed results. In superstring theory, perturbative renormalization can be made to work if it is assumed that the point-like particles we observe — photons, electrons and more exotic subatomic particles — are actually string-like loops that exist in a space–time

with ten dimensions. These ten dimensions include the four dimensions we ordinarily perceive (three of space and one of time) plus six other dimensions that will only show up in experiments at extremely small scales. None of these extra dimensions, or any of the string-like states, have yet been observed.

The two approaches also differ in the way they deal with the most fundamental challenge to the unification of quantum mechanics and general relativity: replacing the fixed space–time arena with a more interesting, dynamical version. In superstring theory the idea of a fixed space–time arena is conserved, but the theory includes some dynamical variables that also describe space–time. The end result is a dynamical space–time, which is really the 'sum' of the classical fixed arena and some dynamical quantum space–time effects. But some physicists feel that the foundation provided by the fixed space–time arena does not fully satisfy the principles of general relativity. In this respect, canonical quantum gravity is more ambitious: space–time is dynamical from the start. Unfortunately it is not yet clear how this dynamical space–time will describe most experimental situations in which space–time does behave as a classical and fixed arena.

A quantum theory of space–time

One of the most exciting possibilities emerging from these and other approaches to the unification of general relativity and quantum mechanics is the idea of a space–time that is itself quantized. This could involve a discretized space–time — for example, replacing the space–time continuum with a collection of isolated points. Or it could involve a quantum 'uncertainty principle' similar to that found in ordinary quantum mechanics. In a space–time with a quantum uncertainty principle it would not be possible to measure accurately the distance between two points in space–time, just as in ordinary quantum mechanics it is not possible to measure simultaneously the position and momentum of a particle within the classical space–time arena.

The quantum description of space–time would require a profound renewal of fundamental physics. For example, it might have unexpected effects on the normal propagation of particles. These effects would only emerge at the small scales of the Planck length (Box 1), which cannot be probed by experiments at present.

A useful analogy is the one of the surface of a wooden table. We usually perceive the surface as perfectly flat, but if we look at the table with a microscope (which allows us to observe its short-distance structure) it becomes clear that it is not flat. The flatness we ordinarily perceive is some sort of averaging over the short-distance irregularities of the table's surface. If we take a ball with diameter much larger than the short-distance scale of roughness of the table and roll it over the surface, we find no evidence of the roughness (Fig. 1a), but if we use a ball whose diameter is not much larger than the scale of roughness, we will see disturbances in the path of the ball resulting from its interaction with the roughness (Fig. 1b).

Similarly, a rich short-distance structure in space–time may have important effects on the propagation of particles, in a way that depends on the wavelengths of the particles. (Because of the wave–particle duality of quantum mechanics, all particles can be characterized by a wavelength, which is inversely related to their momentum.) In the tabletop analogy, the diameter of the balls plays the role of the particle wavelength. For the particles we observe experimentally there should be very little wavelength dependence, but for particle wavelengths comparable with the tiny Planck length, the phenomenon of particle propagation would be severely affected (imagine rolling a ball with diameter smaller than the scale of surface roughness over the table).

Hints of the new revolution

A key ingredient in Planck's revolutionary breakthrough was provided by Ludwig Boltzmann's earlier work, particularly his interpretation of the second law of thermodynamics as a statistical law rather than an absolute law of nature. At the time no one could have imagined what Boltzmann's work would eventually lead to. Is it possible that we already have some of the ingredients of a new revolution today?

In recent years, several theoretical ideas have emerged that would require a change of perspective as radical as Boltzmann's. These include the theories that quantize space–time, discussed above, which explain the classical properties of space–time that are perceived by long-wavelength particles as a sort of average over the richer microscopic structure of quantum space–time. This is not unlike Boltzmann's description of the second law of thermodynamics as a statistical

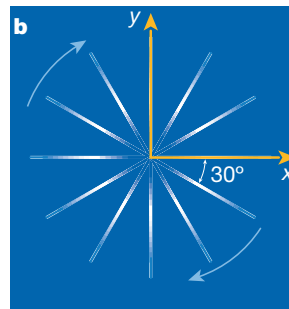
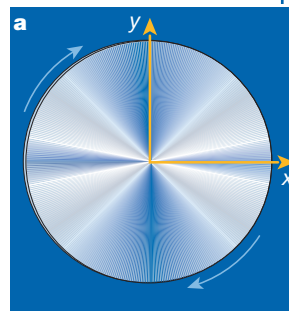
Box 2 Symmetry breaking

Even a space–time that we ordinarily perceive as classically flat and continuous can have a rich quantum structure at short distances (Fig. 1). This has consequences for the fundamental ‘symmetries’ of space–time and for our understanding of particle physics. The idea of a space–time symmetry can be understood by comparing a flat continuous disc (a) and a discretized version of the same disc (b). If the continuous disc (corresponding to a classical space–time) is rotated by any arbitrary angle it will still look the same as the original, but for the discretized disc only certain rotations (integer multiples of 30°) reproduce the original.

Classical space–time is characterized by an important rotational symmetry known as Lorentz symmetry. Particle-physics models that assume a

flat space–time have another fundamental symmetry, known as CPT (the combined symmetries of charge, parity and time reversal), which governs, among other things, the relationship between particles and antiparticles. So far, all the available evidence suggests that the laws of nature obey both Lorentz and CPT symmetry — that is, we only observe the continuous disc. But the low resolution of current experiments cannot distinguish between the continuous and discretized disks — the latter will appear smeared out, as if it were continuous.

One of the key questions for experimental studies of the unification of general relativity and quantum mechanics to answer is the way in which the short-distance quantum structure of space–time affects Lorentz symmetry and CPT symmetry. In other words, if we look at the structure of



space–time with high enough resolution, will we see the discretized disc rather than the continuous disc? **G.A.-C.**

build a case for a quantum description of space–time. Nonetheless, the possibility of being able to test our ideas of quantum gravity with real data is essential for moving this field from the realm of science fiction into science fact.

The equivalence principle

Experiments that can observe the interplay between general relativity and quantum mechanics are not easy to design. The tiny scale of the Planck length (Box 1) at which these effects are expected to emerge, and the extreme weakness of gravity compared with the other forces of nature, have led many to believe that the study of quantum gravity would be beyond the reach of laboratory experiments. Moreover, there appears to be a fundamental theoretical issue that affects the experimental study of gravitational forces at the quantum level. The standard approach to the study of these forces is in conflict with the equivalence principle, one of the cornerstones of general relativity, which says that in a uniform gravitational field all masses fall with the same acceleration (for example, 9.8 m s^{-2} near the surface of the Earth). The equivalence principle is the reason why gravitational forces are so profoundly different from all other forces.

In the study of all other forces, experimental particles used to test the strength of the force field have two primary characteristics: their ‘charge’ under the influence of the force field and their ‘inertial mass’ (defined as the ratio between the force exerted on a body and the resulting acceleration). As shown by Bohr and Rosenfeld¹⁶, the only way to extract accurate information on the force field is to use particles with large inertial mass and small charge. In studies of gravitational fields, this strategy would require particles with large inertial mass and small gravitational mass (which sets the strength of the gravitational forces felt by the particle). But the equivalence principle requires inertial and gravitational masses to be equal. Therefore the standard (Bohr–Rosenfeld) strategy for quantum-mechanics experiments appears to lead^{2,4} to a fundamental limitation on the accuracy with which gravitational fields can be measured.

The experimental frontier

Despite these many challenges, in the last quarter of the twentieth century, physicists have started to make some progress in the experimental study of quantum gravity. The first important step was taken in the mid-1970s with experiments that studied how strong gravitational fields affected the quantum mechanical behaviour of microscopic particles^{17,18}. These experiments were not capable of testing the quantum properties of space–time, but it was established that the gravitational fields generated by the Earth affect ‘interference experiments’ (Box 3) in a

outcome of the properties of the microscopic constituents of a thermodynamic system.

On the experimental side, the crucial input for Planck’s breakthrough came from puzzling data on the radiation emitted by hot objects in thermal equilibrium with their surroundings. The light from these objects is a mixture of different frequencies (colours). The classical formula linking the radiation emitted at different frequencies with the temperature of the hot body works well for radiation at low frequencies, but disagrees with experimental data at high frequencies. Planck observed that this puzzle could be solved by assuming that energy could be emitted or absorbed only in discrete amounts, and the quantum of energy was born.

Among the experimental puzzles confronting physicists today, there are a few that could lead to a profound renewal of fundamental physics. One example is the observation of cosmic rays above the so-called Greisen–Zatsepin–Kuzmin (GZK) limit^{9,10}. Cosmic rays are ultra-high-energy particles (most likely protons) emitted by distant active galaxies. These rays produce showers of elementary particles when they pass through the Earth’s atmosphere. Before reaching us the rays travel gigantic (cosmological) distances, and astronomers would expect interactions with photons of the cosmic microwave background (the faint glow left over from the Big Bang) to cool them

down, producing a cutoff at higher energies. The discovery of cosmic rays above this GZK cut off¹¹, apparently originating from outside our Galaxy, cannot be explained by current theories.

To calculate the GZK limit, theorists have to take into account Lorentz symmetry, a property of classical space–time (Box 2). It has been suggested that the observation of cosmic rays above the GZK limit could be explained by assuming very small modifications to the predictions of Lorentz symmetry^{12,13}. A plausible origin for such a modification comes from an unexpected source — the quantization of space–time. Remarkably, the value of the GZK limit predicted by calculations for some types of quantum space–time turns out to be consistent with the astronomical data¹⁴.

There are other plausible explanations for the cosmic-ray puzzle, but the explanation involving a quantum space–time is the only one that would simultaneously solve another, similar puzzle^{14,15}. Astronomers have detected high-energy photons from distant astrophysical sources that, according to the classical picture of space–time, should not be able to reach us because they are expected to interact with the cosmic infrared background. This means that there should be a theoretical upper limit to the energies of the photons, similar to that for cosmic rays. Experimental evidence for such puzzles is just emerging, so it is too early to

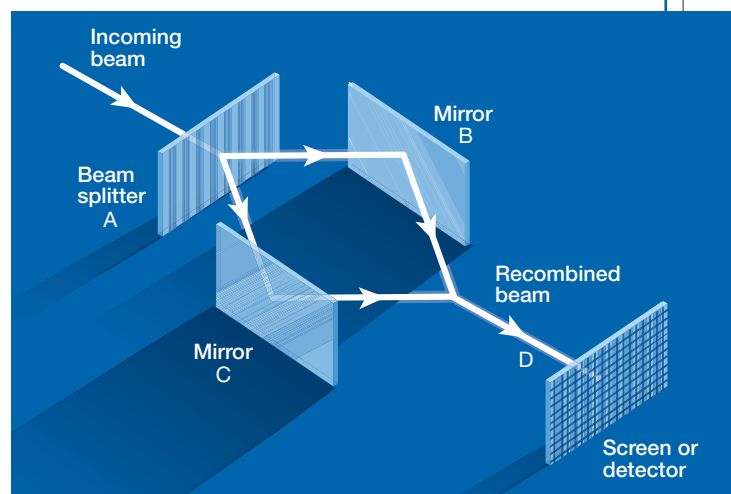
Box 3 Interference experiments

Interference experiments could be sensitive to some of the small effects due to a quantum space–time. In these experiments an important role is played by one of the key features of quantum mechanics: particle–wave duality. Like a wave, a particle beam can be characterized by a wavelength (the separation between the peaks of the wave). Moreover, two different beams can be in phase, if their peaks pass the same points at the same time, or they can have a phase difference. When two beams merge they interfere in a fully constructive way if they are in phase; otherwise the interference is at least partly destructive.

A basic interference experiment is shown here. The incoming

particle beam encounters a beam splitter at A. The two resulting beams are directed towards mirrors at B and C. When they recombine at D the two parts of the beam will be in phase if the two paths (ABD and ACD) are the same length and if they do not encounter physical fields that induce different phase shifts.

One of the predictions of quantum mechanics is that the phase of a wave can be affected by the presence of physical fields, as has been shown for electromagnetic fields. Gravitational fields can also affect the phase difference between the two particle beams if they vary across the apparatus. This is easily achieved by placing the two paths at different heights, producing an observable phase difference. In



certain quantum theories of gravity, the two parts of the beam would experience different quantum

fluctuations of space–time, and this could also give rise to a phase difference.

G.A.-C.

way that is completely analogous to the influence of electromagnetic fields.

In the 1980s, it was realized that experiments testing one of the fundamental symmetries of particle physics^{19–21}, known as CPT (Box 2), were finally reaching sensitivity levels that could plausibly detect minute modifications of these symmetries caused by the quantum properties of space–time. The sensitivity of CPT experiments (in particular studies of particles known as neutral kaons)²² has continued to improve over the past few decades. So far they have provided no evidence of any quantum property of space–time.

At first, tests of CPT symmetry were the only examples of experiments being analysed from the point of view of quantum gravity, but over the past few years more proposals have been put forward. The most sensitive among these experiments are searches for departures from Lorentz symmetry in cosmic-ray physics (Box 2). Besides the GZK limit, discussed above, another key implication of Lorentz symmetry — the prediction that the time needed by a massless particle to propagate over a given distance should not depend on the wavelength of the particle — is going to be tested with extremely high accuracy by astrophysics experiments now in preparation. In particular, a quantum space–time might alter the propagation of γ -rays collected from distant γ -ray bursts — some of the most powerful explosions in the Universe²³. Within 5 or 10 years the next generation of γ -ray telescopes, such as the GLAST space mission²⁴, will reach sensitivity levels sufficient for testing the type of minute wavelength-dependence predicted by certain quantum pictures of space–time.

In addition to tests of Lorentz and CPT symmetries, it might even be possible to explore directly the structure of space–time itself. Modern gravity-wave interferometers (Box 3) are built to detect tiny fluctuations in the distances between test masses that might be caused by a passing gravity wave²⁵. Gravity waves are ripples in the fabric of space–time predicted by Einstein's theory of relativity, but they have never been detected. Last year, I proposed that gravity-wave interferometers could also be used to detect the quantum fluctuations of space–time²⁶. Interferometers under construction, such as the LIGO detector in the United States and the VIRGO detector in Italy, could be sensitive to fluctuations at scales that correspond to the Planck length. So we may get a chance to test directly our fledgling models of quantum gravity, although there is a long way to go before we have a theoretically consistent and experimentally testable picture of a quantum space–time⁴.

Prospects for the coming century

After 100 years of quantum theory and more than 70 years of failed attempts to unify general relativity and quantum mechanics, we finally have some promising theoretical and experimental avenues for the investigation of the interplay between relativity and quantum theory. There is a strong theoretical desire to unite these two theories, but without a healthy experimental programme our best models would remain as mere speculation. For now, the experiments are limited in number and they are only sensitive to effects that might be induced by the unification of general relativity and quantum mechanics at scales larger than the Planck length. Nonetheless, it seems reasonable to expect

that a growing number of experimental and theoretical ideas over the next century will eventually lead to some progress in the unification of quantum mechanics and general relativity.

Giovanni Amelino-Camelia is in the Dipartimento di Fisica, Università 'La Sapienza', I-00185 Roma, Italy.

e-mail: amelino@roma1.infn.it

1. Zeilinger, A. *Nature* **408**, 639–641 (2000).
2. Stachel, J. in *Black Holes, Gravitational Radiation and the Universe: Essays in Honour of C. V. Vishveshwara* (eds Iyer, B. R. & Bhawal, B.) 525–534 (Kluwer Academic, Dordrecht, 1999).
3. Rovelli, C. *J. Math. Phys.* **41**, 3776–3800 (2000).
4. Amelino-Camelia, G. <http://xxx.lanl.gov/abs/gr-qc/9910089>
5. 't Hooft, G. *Quantum Grav.* **16**, 3263–3279 (1999).
6. Ashtekar, A., Rovelli, C. & Smolin, L. *Phys. Rev. Lett.* **69**, 237–240 (1992).
7. Green, M. B., Schwarz, J. H. & Witten, E. *Superstring Theory* (Cambridge Univ. Press, Cambridge, 1987).
8. Polchinski, J. *String theory* (Cambridge Univ. Press, Cambridge, 1998).
9. Greisen, K. *Phys. Rev. Lett.* **16**, 748–751 (1966).
10. Zatsepin, G. T. & Kuzmin, V. A. *Sov. Phys. JETP Lett.* **4**, 78–80 (1966).
11. Bird, D. J. et al. *Astrophys. J.* **441**, 144–151 (1995).
12. Coleman, S. & Glashow, S. L. *Phys. Rev. D* **59**, 116008 (1999).
13. Kifune, T. *Astrophys. J.* **518**, L21–L24 (1999).
14. Amelino-Camelia, G. & Piran, T. <http://xxx.lanl.gov/abs/astro-ph/0008107>
15. Muir, H. *New Sci.* 32–35 (23 September 2000).
16. Bohr, N. & Rosenfeld, L. *Kgl. Danske Videnskab S. Nat. Fys. Medd.* **12**, 1–65 (1933).
17. Colella, R., Overhauser, A. W. & Werner, S. A. *Phys. Rev. Lett.* **34**, 1472–1474 (1975).
18. Ahluwalia, D. V. *Nature* **398**, 199–200 (1999).
19. Ellis, J., Hagelin, J. S., Nanopoulos, D. V. & Srednicki, M. *Nucl. Phys. B* **241**, 381–405 (1984).
20. Kostelecky, V. A. & Potting, R. *Phys. Rev. D* **51**, 3923–3935 (1995).
21. Huet, P. & Peskin, M. E. *Nucl. Phys. B* **434**, 3–38 (1995).
22. Ellis, J., Lopez, J., Mavromatos, N. E. & Nanopoulos, D. V. and CPLEAR Collaboration, *Phys. Lett. B* **364**, 239–245 (1995).
23. Amelino-Camelia, G., Ellis, J., Mavromatos, N. E., Nanopoulos, D. V. & Sarkar, S. *Nature* **393**, 763–765 (1998).
24. de Angelis, A. <http://arXiv.org/abs/astro-ph/0009271>
25. Saulson, P. R. *Fundamentals of Interferometric Gravitational Wave Detectors* (World Scientific, Singapore, 1994).
26. Amelino-Camelia, G. *Nature* **398**, 216–218 (1999).

A giant atomic slide-puzzle

Atoms crammed tightly together in metal crystal surfaces are surprisingly mobile.

Atoms in close-packed surfaces of metal crystals move around at surprisingly high rates, even though each atom is locked in tightly by its neighbours. Here we use a low density of indium atoms, embedded in the outermost atomic layer of a copper surface, as tracer particles for scanning tunnelling microscopy to reveal the high vacancy-assisted mobility of atoms in this surface. We believe that most close-packed surfaces of metals and other materials will exhibit a similar vacancy-assisted motion at room temperature, with such surfaces behaving like a gigantic atomic slide-puzzle.

We deposited indium on a clean Cu(001) surface, one of the naturally occurring facets on a copper crystal, in an ultrahigh vacuum system and then watched the diffusion of individual indium atoms at room temperature using a scanning tunnelling microscope¹ (STM). The surface consists of atomic 'steps' separating atomically flat copper 'terraces'. The indium atoms are rapidly incorporated in the outermost atomic layer, where each one replaces a single copper atom (refs 2,3; and R.v.G. *et al.*, manuscript submitted).

To our surprise, we found that the embedded indium atoms are mobile and that they 'jump' over distances larger than one lattice spacing, some of them as far as five lattice spacings. The root-mean-square jump length is 1.92 lattice spacings; these jumps are separated by relatively long intervals of about 100 seconds at room temperature. Neighbouring indium atoms show a strong tendency to jump simultaneously (Fig. 1 represents a typical sequence of STM images illustrating these observations).

We can explain this peculiar motion of the indium by assuming that it is assisted by a rapidly diffusing 'mystery particle' that remains invisible to the STM. During its fast two-dimensional random walk, there will be a high probability that this particle will encounter an indium atom several times, thereby displacing it over more than a single lattice spacing in a time that is too short to be resolved by the STM and giving rise to the long jumps. The mystery particle will also have a high probability of encountering other indium atoms in the direct vicinity, which explains the simultaneous jumps.

There are two naturally occurring particles that might act in the assisting role invoked here, namely adatoms (single copper atoms on top of the surface) and vacancies (atoms missing from the first layer). We have observed that the indium atoms, immediately after deposition on top of the surface, invade the first copper layer via the steps, which rules out the possibility that

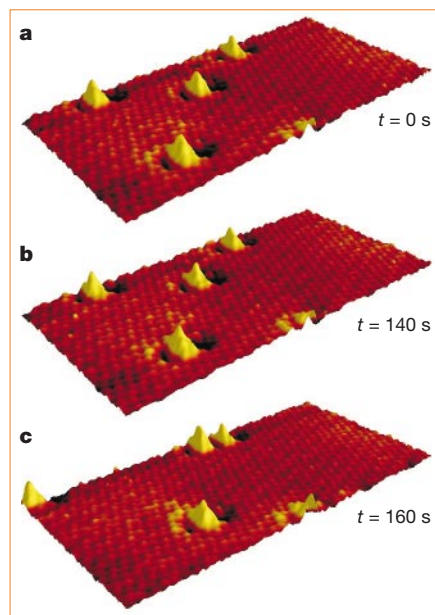


Figure 1 Images ($14 \times 7 \text{ nm}^2$) taken from a movie recorded by scanning tunnelling microscopy (see Supplementary Information), illustrating the diffusion of embedded indium atoms at room temperature. The fine corrugation is due to the atomic lattice of the Cu(001) surface. The four protrusions are individual indium atoms, each one replacing a single copper atom in the surface layer. **a**, Starting situation at $t = 0 \text{ s}$; **b**, the positions of the four indium atoms are still identical to those in **a**, despite the fact that 140 s has passed; **c**, 20 s later, three of the four indium atoms have moved up to 5 atomic spacings. All motion exhibits this pattern of long waiting times followed by simultaneous displacements of several indium atoms over long distances.

adatoms assist the indium motion by changing places with the indium — in that case, most indium atoms would have entered the copper layer at a point close to where they had landed, resulting in a homogeneous population of the terraces with indium. We conclude that surface vacancies are responsible for the diffusion of the indium atoms.

Our interpretation that each long jump reflects the passage of a single assisting entity — in this case a vacancy — is corroborated by the measured distribution of jump lengths. The shape of this distribution is described well by a modified Bessel function, as expected for this type of diffusion (ref. 4; and R.v.G. *et al.*, manuscript submitted), rather than the ordinary Gauss function expected for an unassisted random walk of the indium. The distribution of waiting times between successive jumps is purely exponential, which shows that subsequent (long) jumps are the effect of the passage of different vacancies.

The indium atoms should be regarded merely as tracer particles — the observed motion of the indium reflects the diffusion of all copper atoms in the surface layer. Model calculations show, however, that the width of the measured jump-length distribution implies a noticeable short-range attraction between the indium and the vacancies, which makes the average jump length of the indium atoms somewhat larger than that of the copper atoms (without affecting the average jump frequency). The mechanism by which surface vacancies allow the diffusion of atoms in the surface is like a slide-puzzle, in which a set of square tiles can be rearranged completely by sliding a single missing tile through the puzzle.

R. van Gastel*, E. Somfai†,

W. van Saarloos†, J. W. M. Frenken*

*Kamerlingh Onnes Laboratory, PO Box 9504, and
†Instituut-Lorentz, PO Box 9506, Universiteit
Leiden, 2300 RA Leiden, The Netherlands

1. Hoogeman, M. S. *et al.* *Rev. Sci. Instr.* **69**, 2072–2080 (1998).
2. Breeman, M. & Boerma, D. O. *Phys. Rev. B* **46**, 1703–1709 (1992).
3. Breeman, M. & Boerma, D. O. *Surf. Sci.* **287/288**, 881–885 (1993).
4. Brummelhuis, M. J. A. M. & Hilhorst, H. J. *J. Stat. Phys.* **53**, 249–278 (1988).

Supplementary information is available on Nature's World-Web site (<http://www.nature.com>).

Psychology

An electoral butterfly effect

Part of the controversy surrounding this year's presidential election in the United States concerns the potential for systematic bias in the ballot-card format — could the butterfly ballot used in Palm Beach County, Florida, have led to confusion and caused people who had intended to vote for Al Gore to vote for Pat Buchanan by mistake? Here we show that not only is the double-column butterfly ballot more con-

fusing than a single-column ballot, but that it also appears to cause systematic errors in voting which call into question the validity of the results from Palm Beach County in the 2000 United States presidential election.

To test whether the butterfly ballot format is likely to confuse voters, we asked Canadian college students to vote for a prime minister of Canada on the day after the presidential election in the United States, using either a single-column ballot or a dual-column (butterfly) ballot like the one used in Palm Beach County (fortuitously, there was a Canadian federal election on 27 November 2000).

The ballots contained the names of leaders of ten Canadian political parties and space for a write-in candidate. The dual-column butterfly ballot was designed so that the leaders of the two dominant parties appeared in the first and second positions in the first column (Fig. 1). Stockwell Day, leader of the Canadian Alliance Party, was in the first position, corresponding to the position occupied by George W. Bush's name in the Palm Beach County butterfly ballot, and Jean Chrétien, leader of the Liberal Party, was in the second position, corresponding to the position of Gore's name. The leader of another party, expected to receive few votes, was the first name to appear in the second column. Specifically, Joe Clark, leader of the Progressive Conservative Party, was in the position corresponding to that of Buchanan in the Palm Beach County ballot. The remaining candidates were from parties that were expected to receive only a few votes.

Participants ($n = 324$) randomly received one of the two types of ballot (single- or double-column) and voted for a prime minister by darkening the circle beside their preferred candidate's name; they then evaluated any confusion caused by the ballot format (calculated using the mean, M , of two items on 7-point scales, with high scores indicating greater confusion; Cronbach's $\alpha = 0.96$) and reported the name for whom they had intended to vote.

The results showed that the butterfly ballot ($M = 3.69$, $n = 161$) was more confusing than the single-column format ($M = 2.14$, $n = 163$; $t(322) = 8.23$, $P < 0.0001$); however, none of the students made any errors. Although greater confusion might be expected to lead to higher error rates, we were not surprised by the lack of error in this sample because it involved students skilled at completing complex scoring sheets. We therefore decided to collect data off campus from a sample more representative of the general population.

The design of these ballots was similar to the ones used by the students, except that by 9 November 2000 we were able to modify the butterfly ballot to resemble exactly the format used in Palm Beach County (but without punch holes; Fig. 1). Participants (51 males, 62 females and 3 respondents who failed to report their gender; mean age was 51.10 years, s.d. = 19.19; range, 19–86 years) were approached individually at a mock polling station in the Bonnie Doon Shopping Centre in Edmonton, Alberta, and were randomly assigned to one of the two ballots; they were then directed to one of two polling booths to vote for a prime minister. We subsequently asked the participants to evaluate ballot confusion in the same way as the student sample had (Cronbach's $\alpha = 0.82$), to report for whom they had intended to vote, and to give their gen-

(CANADIAN ALLIANCE)			
STOCKWELL DAY	3→	⊙	(PROGRESSIVE CONSERVATIVE)
(LIBERAL)		⊙	←4 JOE CLARK
JEAN CHRETIEN	5→	⊙	(COMMUNIST)
(NEW DEMOCRATIC)		⊙	←6 MIGUEL FIGUEROA
ALEXA MCDONOUGH	7→	⊙	(BLOC QUEBECOIS)
(GREEN)		⊙	←8 GILLES DUCEPPE
JOAN RUSSOW	9→	⊙	(CANADIAN ACTION)
(NATURAL LAW)		⊙	←10 PAUL HELLIER
NEIL PATERSON	11→	⊙	WRITE IN CANDIDATE
(LIBERTARIAN)		⊙	
JEAN-S BRISSON	13→	⊙	

Figure 1 The arrangement of candidates on the butterfly ballot in a mock election for the Canadian prime minister. The card shows the entry made by an individual who intended to cast a vote for Chrétien and instead erroneously cast a vote for Clark (this error is directly comparable to a Gore–Buchanan voting mistake because their names occupied similar positions on the butterfly ballot). Note that the design of the ballot could lead a person to commit such an error because Chrétien is in the second position in the first column, yet the circle corresponding to Chrétien's name is in the third position on the ballot; this type of error would appear to be less likely for Day (or Bush on the US ballot).

der, age and ethnic background, and finally to place their ballots in a ballot box.

Four people failed to complete both confusion items. The results indicated that the butterfly ballot was more confusing ($M = 3.52$, $n = 53$) than the single-column ballot ($M = 2.30$, $n = 59$; $t(110) = 3.32$, $P < 0.002$). There were four errors, all of which occurred on the butterfly ballot (that is, a 7.55% (4/53) error rate on the butterfly ballot compared with 0% on the single-column ballot; likelihood ratio $\chi^2(1) = 5.47$, $P < 0.02$). Three of the four errors occurred against the candidate who occupied the same position on our butterfly ballot as Gore on the Palm Beach County ballot — this candidate's votes were unintentionally given to the candidate who was in the same position as presidential candidate Buchanan (that is, of the 15 people who intended to vote for Chrétien, 20% erroneously voted for Clark instead, which corresponds to a Gore–Buchanan error on the Palm Beach County ballot). Thus, the butterfly ballot appears to cause systematic errors in the casting of votes.

Considering only those participants exposed to the butterfly ballot, there was no relation between age and errors ($r = 0.05$, n.s.), or between the amount of confusion and errors ($r = 0.07$, n.s.). Furthermore, participants who made errors ($M = 4.00$, $n = 4$) did not differ in their confusion from those who did not commit errors ($M = 3.45$, $n = 49$; $t(51) = 0.53$, n.s.).

We asked participants after voting whether they were aware of the butterfly

ballot controversy in Palm Beach County, as it might have influenced their responses. In the student study, no one was suspicious of the ballot format; in the second study, only three voters were hypothesis suspicious and none of these made errors. If anything, then, it appears that awareness of the ballot issue was associated with careful voting.

The results from these two studies indicate that the butterfly ballot is more confusing than a single-column ballot. Our study on the second group of voters suggests that the butterfly ballot may cause systematic errors in voting which could cast doubt on the validity of the results from the Palm Beach County vote. With the butterfly ballot, vote counts systematically vary from the intention of the electorate.

It is unclear whether a biasing ballot format does or should have legal standing in adjudicating disputes after an election. But given the centrality of elections to the democratic process, it is remarkable that biasing formats continue to be used. Low-cost application of social science theory and methods would help to prevent such controversy in the future.

Robert C. Sinclair, Melvin M. Mark*, Sean E. Moore, Carrie A. Lavis, Alexander S. Soldat

Department of Psychology, P-343 BioSci, University of Alberta, Edmonton, Alberta T6G 2E9, Canada
e-mail: sinclair@ualberta.ca

**Department of Psychology, Pennsylvania State University, Moore Building, University Park, Pennsylvania 16802, USA*

Geophysics

Timing of the Martian dynamo

On Mars, the strong magnetization in the highland crust of the southern hemisphere and the absence of magnetic anomalies at the Hellas and Argyre impact basins have been taken as signs that the core dynamo that once drove the planet's magnetic field turned off more than 4 billion years (Gyr) ago. Here, we argue instead that the Martian dynamo turned on less than 4 Gyr ago and turned off at an unknown time since then. High spatial resolution magnetometry in both Martian hemispheres is needed to reveal the true history of the Martian dynamo.

The discovery by Mars Global Surveyor of remanent crustal magnetization was strong evidence that Mars — which now has no magnetic field — once had a core dynamo¹. The onset and duration of dynamo action place strong constraints on a planet's thermal evolution. The persistence of the Earth's dynamo for the past 3 Gyr is attributed to the solidifying of the

inner core^{2,3}, with the consequent release of latent heat and with gravitational energy powering the magnetic field.

Terrestrial planets with a liquid core and no solidification of the inner core may lack a dynamo — as proposed for Venus^{4,5}. Although vigorous thermal convection in a liquid core could theoretically generate a magnetic field⁴, no examples of this have been found in the terrestrial planets so far. Therefore, although an early dynamo driven by the cooling of a hot liquid core is possible, the most likely scenario for a terrestrial-type dynamo is onset after the beginning of inner-core solidification and shut-off when the core is substantially frozen.

The Moon provides support for this hypothesis. Correlation of Apollo subsatellite magnetometer data with lunar geology shows that magnetic fields were stronger over Imbrian age units than pre-Imbrian, consistent with a late dynamo turn-on⁶. Lunar palaeointensity data show that a dynamo turned on relatively abruptly about 4 Gyr ago and that the magnetic field became weaker over 1 billion years⁷. This late onset of the lunar dynamo may mark the beginning of inner-core solidification.

The belief that the Martian dynamo stopped after only several hundred million years has led to theories such as the cessation of a plate-tectonic style of mantle convection more than 4 Gyr ago⁸. But we are not convinced that the early dynamo interpretation is correct, and believe the onset time of the Martian dynamo is uncertain.

Rather than requiring a dynamo that turned on and off over 4 Gyr ago, the evidence suggests that the dynamo did not begin until well after this. Only the weakness of the present Martian magnetic field limits its duration. The absence of magnetic anomalies at the Hellas and Argyre basins implies that the Martian dynamo did not exist until after the bulk of the southern hemisphere's crust had formed. If it was operative then, the crust should have been magnetized. Large impacts that subsequently punched holes in the crust would have produced distinctive magnetic anomalies. As no anomalies associated with impact basins have been observed, the bulk of the crust in the south is not magnetized and there was no Martian dynamo at crustal formation or when the basins formed at about 4 Gyr or earlier. Impacts into the previously unmagnetized crust of a Mars with a magnetic field should have created magnetic anomalies.

Although we cannot rule out the possibility that the dynamo turned on after the southern crust formed and stopped before the major impact basins were formed, it is easier to explain its turn-on after 500 Myr of core evolution than its turn-off before this time. Models of core cooling suggest that a dynamo lasts longer than a few hun-

dred million years, especially as core cooling leads to inner-core solidification that powers the dynamo as long as the inner core continues to grow^{4,9}. This dynamo action might await the onset of inner-core growth, which could take 500 Myr or more. So although dynamo action could be shortened by a change in Martian mantle convection from a plate-tectonic regime (efficient core cooling) to a rigid-lid regime (inefficient core cooling)⁸, we still believe that the onset did not occur until after 500 Myr of core evolution and the formation of the major impact basins.

If the dynamo turned on after the giant impact basins formed, then the magnetized southern regions must either be later magmatic additions to the crust (after 500 Myr of evolution) or thermal reworking of older crust. Although, on average, the crust has cooled over time, local regions have been heated by upwelling plumes. The non-uniformity of the magnetization in the south suggests that it arose mainly from localized heating and cooling events that postdate the global cooling to below the Curie point of the southern highland crust. As there does not seem to be magmatic activity in most of the magnetized regions, these heating events must have been due to upwellings smaller than those responsible for Tharsis and the Elysium volcanoes.

There is no unambiguous constraint on the dynamo turn-off time other than its absence at present. The remanence of the SNC (Shergotty, Nakhla and Chassigny) meteorites with formation ages of 1.3 Gyr to 180 Myr is consistent with ancient surface fields of 500 to 5,000 nanotesla (ref. 10), but there could be such fields at the surface of Mars even today. Nevertheless, the presence of localized magnetic anomalies in the younger crust of the northern hemisphere¹¹ indicates that the dynamo could not have turned off too quickly. Strongly magnetized terrain also extends from the southern hemisphere highlands into the Tharsis region, one of the youngest surfaces on Mars, emplaced well after the heavy bombardment¹².

We believe that high spatial resolution magnetometry with balloons or airplanes in both northern and southern hemispheres would resolve the nature and timing of the magnetization, much as shipborne magnetometers improved our understanding of the Earth's dynamo and plate tectonics.

G. Schubert, C. T. Russell, W. B. Moore

*Department of Earth and Space Sciences, and
Institute of Geophysics and Planetary Physics,
University of California, Los Angeles,
California 90095-1567, USA*

e-mail: schubert@ucla.edu

1. Acuña, M. H. *et al.* *Science* **284**, 790–793 (1999).

2. Loper, D. E. *J. Geophys. Res.* **83**, 5961–5970 (1978).

3. Gubbins, D. *et al.* *Geophys. J. R. Astron. Soc.* **59**, 57–99 (1979).

4. Stevenson, D. J., Spohn, T. & Schubert, G. *Icarus* **54**, 466–489 (1983).

5. Stevenson, D. J. *Rep. Prog. Phys.* **46**, 555–620 (1983).

6. Russell, C. T. *et al.* in *Proc. 8th Lunar Sci. Conf.* 1171–1185 (Pergamon, New York, 1977).

7. Runcorn, S. K. *Phil. Trans. R. Soc. Lond. A* **349**, 181–196 (1994).

8. Nimmo, F. & Stevenson, D. J. *Geophys. Res.* **105**, 11969–11980 (2000).

9. Schubert, G. *et al.* in *Mars* (eds Kieffer, H. H. *et al.*) 147–183 (Univ. Arizona Press, Tucson, 1992).

10. Collinson, D. W. *Meteor. Planet. Sci.* **32**, 803 (1997).

11. Purucker, M. *et al.* *Geophys. Res. Lett.* **27**, 2449 (2000).

12. Strom, R. G., Croft, S. K. & Barlow, N. G. in *Mars* (eds Kieffer, H. H. *et al.*) 383–423 (Univ. Arizona Press, Tucson, 1992).

ATP-dependent proteases

Docking of components in a bacterial complex

Proteases are enzymes that cut up other proteins for the purposes of tailoring or degradation. Some depend on ATP as an energy source for unfolding protein substrates, and these are often organized into rings of ATPase subunits stacked coaxially onto rings of protease subunits¹. Bochtler *et al.*² have reported a crystal structure for the ATP-dependent protease complex HslVU (also known as ClpYQ) from *Escherichia coli*. They claim this consists of a double hexamer of the protease HslV flanked by hexamers of an ATPase, HslU, which mainly lie in a ring of ATPase domains whose I-domains protrude to form a smaller ring that binds HslV. Based on cryo-electron microscopy of HslVU in buffer conditions that support enzymatic activity, we find that the HslU rings bind in the opposite orientation — that is, their I-domains protrude distally instead of making contact with HslV. Redefinition of this interaction has implications for the functional architecture of the complex.

The components of HslVU² share features with other protein structures: HslV is similar in sequence and fold to the proteasome β -subunit³, and the ATPase domain of HslU resembles the D2 domain of the protein NSF (for N-ethylmaleimide-sensitive fusion protein)⁴, apart from containing an insertion of a 133-residue intermediate (I) domain. In the reported crystal structure of HslVU², the I-domain ring is in contact with HslV. This configuration is inverted relative to that suggested by negatively stained electron micrographs, which show the wider, denser ring of HslU in contact with HslV⁵. Bochtler *et al.*² attribute these earlier observations to a flattening artefact, and expand on the functional implications of the contacts between the I-domain and the protease.

We observed HslVU preserved virtually in its native state by cryo-electron microscopy. The averaged side view at 30 Å resolution agrees well with our negatively stained rendition of the complex (compare Fig. 1a to b, c): in both cases, the wider, larger ring of HslU is adjacent to HslV. In this respect,

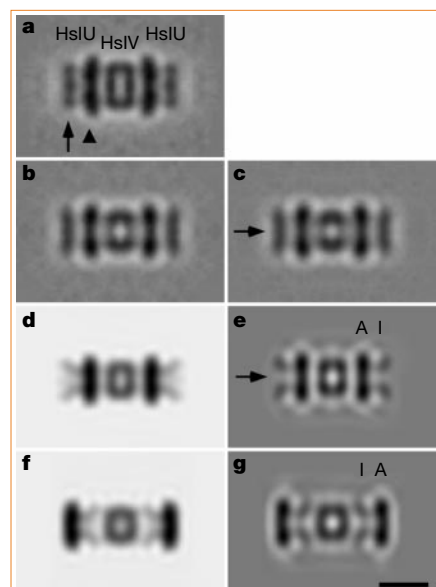


Figure 1 Averaged side-view projections of HslVU complexes. **a–c**, Electron micrographs of complexes formed in 50 mM Tris–HCl, pH 7.5, 0.2 M KCl, 10 mM MgCl₂ and 1 mM ATP, representing appropriate conditions for proteolytic activity. **a**, Negatively stained molecules (ATP-γS state; number of particles, $N=65$; resolution, 32 Å). Note that the proximal ring of HslU (arrowhead) is wider and more dense than the outer ring (arrow); **b**, frozen-hydrated molecules (ATP-γS state; $N=250$; resolution, 33 Å); **c**, frozen-hydrated molecules in the AMP-PNP state, where AMP-PNP is an inactive ATP analogue ($N=400$; resolution, 33 Å). **d–g**, Side-view projections calculated from the crystal structure² but limited to 30 Å resolution. Projections corresponding to different rotational settings of the complex around the axis were averaged to give a cylindrically averaged side view, as in the electron micrographs (EMs). In **d** and **e**, HslU is in the opposite orientation from the one in the crystal structure, whereas in **f** and **g** this corresponds to the published orientation². Projections shown in **e** and **g** were created by applying a phase-contrast transfer function (CTF; corresponding to 2.0 μm underfocus) to images in **d** and **f**, and so are more comparable to the cryo-EMs. With or without CTF correction, it is evident that the wider, denser ring, corresponding to the ATPase domains of HslU, is adjacent to HslV. Arrows in **e** and **c** mark the axial density that is missing in **e** but present in **b** and **c**, which we attribute to residues 175 to 209. In **e** and **g**, I denotes the I-domain ring, and A denotes the ATPase domain ring. Scale bar, 100 Å.

the images are inconsistent with a side-view projection (Fig. 1f) generated from the published coordinates of HslVU² and limited to the same resolution⁶.

For closer comparison, we subjected this image computationally to phase-contrast effects (Fig. 1g), simulating those of the cryo-electron microscopy images. Conversely, an excellent match was obtained with similarly generated projections (Fig. 1d, e) in which HslU was inverted (compare Fig. 1c, e). We conclude that the I-domains are exposed on the distal surfaces of the HslVU complex, and the opposite face of the HslU ring binds to HslV. Despite good overall agreement with the results from cryo-electron microscopy, the calculated re-projection shows the central part of the distal ring of HslU as relatively depleted in density (arrows in Fig. 1c, e). We infer that

the additional density in the electron microscopy images represents six copies of residues 175–209 which were not seen in Bochtler *et al.*'s crystal structure².

HslU belongs to the AAA superfamily⁷ of ATPases, as do the ATPases of the 26S proteasome, and ClpA and ClpX of *E. coli*, which both partner the protease ClpP (ref. 1). As demonstrated for ClpA⁸ and ClpX⁹, all such ATPases are likely to have protein 'unfoldase' activity. Processive degradation is carried out by fully assembled holo-enzymes¹⁰ and requires the coordinated activity of multiple sites. The geometry of interaction between the ATPase and proteinase rings is crucial in specifying the positions of the sites at which substrates bind, where they are unfolded, and the path along which they translocate into the digestion chamber inside the protease.

Our model assigns the I-domains to the distal surfaces of the HslVU complex, in an exposed position that would be suitable for substrate binding. This proposition is consistent with data showing protein substrates binding to the distal surfaces of both ClpXP¹¹ and ClpAP (T.I. *et al.*, manuscript submitted). In this revised model, residues in the ATPase domain of HslU, which includes the carboxy-terminal sensor-2 domain⁷, are responsible for binding to HslV.

Takashi Ishikawa, Michael R. Maurizi*, David Belnap, Alasdair C. Steven

Laboratory of Structural Biology, National Institute of Arthritis, Musculoskeletal and Skin Diseases and *Laboratory of Cell Biology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

e-mail: steven@calvin.niams.nih.gov

- Schmidt, M., Lupas, A. N. & Finley, D. *Curr. Opin. Chem. Biol.* **3**, 584–591 (1999).
- Bochtler, M. *et al.* *Nature* **403**, 800–805 (2000).
- Lowe, J. *et al.* *Science* **268**, 533–539 (1995).
- Yu, R. C., Hanson, P. L., Jahn, R. & Brunger, A. *Nature Struct. Biol.* **5**, 803–811 (1998).
- Rohrwild, M. *et al.* *Nature Struct. Biol.* **4**, 133–139 (1997).
- Belnap, D. M. *et al.* *J. Struct. Biol.* **25**, 166–175 (1999).
- Neuwald, A. F. *et al.* *Genome Res.* **9**, 27–43 (1999).
- Weber-Ban, E. U., Reid, B. G., Miranker, A. D. & Horowitz, A. L. *Nature* **401**, 90–93 (1999).
- Kim, Y. I. *et al.* *Mol. Cell* **5**, 639–48 (2000).
- Singh, S. K., Guo, F. & Maurizi, M. R. *Biochemistry* **38**, 14906–14915 (1999).
- Ortega, J., Singh, S. K., Ishikawa, T., Maurizi, M. R. & Steven, A. C. *Mol. Cell* (in the press).

Bochtler et al. reply — The central issue raised by Ishikawa *et al.* is that of the configuration of the productive HslVU complex. We note that complexes of HslV and HslU in *E. coli* are labile and unstable under many conditions¹. The original electron microscopy (EM) images of Rohrwild *et al.*² appear to show free HslV, free HslU and HslV–HslU complex particles. To explain the discrepancy between the HslV–HslU arrangement in our co-crystals and their negatively stained EM data, we suggested that there might have been a collapse of the fragile I-domain structure in the EM prepa-

rations, or a reversal in the orientation of the HslU rings¹. Ishikawa *et al.* interpret their results from cryo-EM at 30 Å resolution in this latter way, using our crystal data of the components. Although these preparations preserve the native structure better than the negatively stained ones, our HslV–HslU samples are also active under crystallization conditions³.

We have attempted to distinguish between the two docking modes (I-domains distal or proximal to HslV) in mutagenesis experiments involving more than two dozen mutants³. We disrupted putative contact sites to HslV in the I-domain of HslU (Fig. 3b in ref. 1) and on its opposite face and find none of these mutations has any effect on peptide hydrolysis or on casein degradation. This suggests either that no precise complex is required, or that both modes of docking are feasible. In contrast, degradation of the physiological substrate fusion protein MBP–SulA is affected by mutations both in the I-domain as well as those involving the opposite side of HslU. Small-angle X-ray scattering data and a crystal structure of the *Haemophilus influenzae* HslVU complex are also consistent with the EM docking mode⁴.

Matthias Bochtler*, Claudia Hartmann, Hyun Kyu Song, Ravishankar Ramachandran, Robert Huber

Abteilung Strukturforschung, Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, Planegg-Martinsried D-82152, Germany e-mail: huber@biochem.mpg.de

*International Institute of Molecular and Cell Biology, ul. Ks. J. Trojdena 4, 02-109 Warszawa, Poland

- Bochtler, M. *et al.* *Nature* **403**, 800–805 (2000).
- Rohrwild, M. *et al.* *Nature Struct. Biol.* **4**, 133–139 (1997).
- Song, H. K. *et al.* *Proc. Natl Acad. Sci. USA* (in the press).
- Sousa, M. C. *et al.* *Cell* **103**, 633–643 (2000).

Addendum

A new model for protein stereospecificity

A. D. Mesecar & D. E. Koshland Jr

Nature **403**, 614–615 (2000)

To clarify possible misunderstanding over the term "new" in this communication, we meant our new model replaces the old Ogston model, as this is incompatible with our X-ray crystallographic data. We did not intend "new" in this context to mean that nobody had ever questioned the Ogston model before (for example, see refs 1–5).

- Wilcox, P. E., Heidelberger, C. & Potter, V. R. Chemical preparation of asymmetrically labeled citric acid. *J. Am. Chem. Soc.* **72**, 5019–5024 (1950).
- Bentley, R. Ogston and the development of the prochirality theory. *Nature* **276**, 673–676 (1978).
- Bentley, R. Three point attachment: past, present and no future. *Ann. NY Acad. Sci.* **41**, 5–24 (1983).
- Topiol, S. & Sabio, M. Interactions between eight centres are required for chiral recognition. *J. Am. Chem. Soc.* **111**, 4109–4110 (1989).
- Temussi, P. A. *et al.* Structure activity relationship of sweet molecules. *Am. Chem. Soc. Symp. Ser.* **450**, 143–161 (1990).

Editorial note: As Brief Communications do not allow space for conventional introductions, we asked the authors for this addendum to clarify a possible misinterpretation.

Granite magma formation, transport and emplacement in the Earth's crust

N. Petford*, A. R. Cruden†, K. J. W. McCaffrey‡ & J.-L. Vigneresse§

* Centre for Earth and Environmental Science Research, Kingston University, Surrey KT1 2EE, UK

† Department of Geology, University of Toronto, Toronto, Ontario M5S 3B1, Canada

‡ Department of Geological Sciences, University of Durham, Durham, DH1 3LE, UK

§ CREGU, UMR 7566 CNRS, BP 23, 54501 Vandoeuvre Cedex, France

The origin of granites was once a question solely for petrologists and geochemists. But in recent years a consensus has emerged that recognizes the essential role of deformation in the segregation, transport and emplacement of silica-rich melts in the continental crust. Accepted petrological models are being questioned, either because they require unrealistic rheological behaviours of rocks and magmas, or because they do not satisfactorily explain the available structural or geophysical data. Provided flow is continuous, mechanical considerations suggest that—far from being geologically sluggish—granite magmatism is a rapid, dynamic process operating at timescales of $\leq 100,000$ years, irrespective of tectonic setting.

The Earth's granite crust harbours continental landmasses that have remained stable and above mean sea level for more than a billion years¹. The withdrawal of large volumes of granitic magma from the deep continental crust and its emplacement at higher structural levels produces a dehydrated and refractory lower crust, and an upper crust enriched in felsic (iron- and silica-rich) minerals and heat-producing (radioactive) elements. This flux of siliceous melts from deeper levels has also concentrated many chemical elements important for life (including potassium, sodium, phosphorus, lithium and chlorine) close to the surface.

Continental granite magmatism involves four separate but potentially quantifiable stages—generation, segregation, ascent and emplacement^{2,3}—that operate over length scales ranging from 10^{-5} to 10^6 m. The nature of granite magmatism and associated continental growth has been investigated mostly through geochemical and isotope studies. But since the early 1990s, research into the origin of granite has shifted away from geochemistry towards understanding the underlying physical processes involved. As a result, dynamic models that operate on timescales of months to centuries are replacing the once-prevailing view of granitic magma production as a slow, equilibrium process that requires millions of years for completion.

Here we review the important physical processes that control how granitic melts are extracted from their source regions, transported through the Earth's crust and intruded (often with cumulative volumes in excess of 4.5×10^5 km³) into pre-existing rock. Most granite plutons found in the upper continental crust seem to be emplaced as low-viscosity, crystal-poor ($\leq 30\%$) magmas in tabular intrusions fed from depth by small magma batches that ascend rapidly, either in relatively thin conduits or channelled along shear zones. Space for the incoming magma is made by a combination of lateral and vertical displacements at moderate strain rates (typically $>10^{-14}$ s⁻¹) on timescales of less than 100,000 years.

Partial melting of continental crust

Given the average thickness of the continental crust of about 35 km, typical geothermal gradients of 20°C km^{-1} do not generate temperatures high enough to melt common crustal rocks (generally more than 800°C)⁴. Although minor volumes ($<25\%$) of partial melt can be produced by fluid-present melting at high water fugacity ($a_{\text{H}_2\text{O}}$) during thermal relaxation of thickened orogens⁵, local melting is far more efficient where heat is advected into the crust from the underlying (hotter) mantle by basaltic magmas^{6,7}.

Partial melting of crustal rocks pre-heated in this way is likely to be rapid, with models predicting a melt layer two-thirds the thickness of the basaltic intrusion forming in 200 years^{4,6}, at a temperature of up to 950°C . Recent experiments on natural rock systems at temperatures comparable with those attained during basaltic underplating have shown the importance of reactions involving the breakdown of micas and amphiboles to produce melts of granitic composition under fluid-absent conditions (see refs 4 and 8 for a review). These reactions are characterized by steep, positive pressure–temperature slopes with most hydrous phases breaking down over a narrow temperature interval (about $50\text{--}80^\circ\text{C}$), giving rise to melt fractions in the range 0.2–0.4 (Fig. 1). Compositional differences are reflected in higher melting temperatures⁹, with metapelites yielding 20% melt at around 800°C , whereas amphibolites require temperatures close to 950°C to generate similar amounts of melt. An important consequence of fluid-absent melting is the moderate to large positive volume change that accompanies some reactions¹⁰. For the case of crust heated from below, reaction-induced volume changes will lead to local fracturing and an accompanying reduction in rock strength caused by increased pore fluid (melt) pressures^{8,11}. Deviatoric stress gradients can also develop in the vicinity of the intruding mafic heat source and promote local fracturing. These processes, in conjunction with regional tectonic strain, are important in providing enhanced fracture permeabilities of about $10^{-10}\text{--}10^{-5}$ m² in the region of partial melting¹², which aid subsequent melt segregation.

Melt transport

Melt and magma (melt plus suspended solids) must be transported both within and out of the source region before final emplacement. The transport process operates on two length scales: segregation, marked by small-scale movement of melt (centimetres to decimetres), mostly within the source region, and long-range (kilometre-scale) ascent through the continental crust to the site of final emplacement.

Segregation. The ability of granitic melt to segregate mechanically from its matrix is strongly dependent on its physical properties, of which viscosity and density are the most important. The viscosity of silicate melt is a function of composition, temperature and water content¹³. For much of this century, granitic melt was commonly thought to have a viscosity close to that of solid rock¹⁴. However, recent experimental studies have found that granitic melts generated over a wide range of crustal pressure–temperature conditions have viscosities generally in the range $10^8\text{--}10^3$ Pa s (refs

13–15). Figure 2 shows estimates of average melt viscosities calculated using models reviewed in ref. 15 for a typical tonalite (SiO_2 , 65 wt%) and leucogranite (SiO_2 , 75 wt%) as a function of the water content of the melt. The median viscosity (η) of both liquids at their respective ‘ideal’ water contents of 4 and 6 wt% respectively is $10^{3.8}$ and $10^{4.9}$ Pa s, a difference of less than about 1.1 $\log_{10}$ units. An important implication of this material similarity is that those aspects of the segregation and ascent process moderated by the physical properties of the liquid should occur at broadly similar rates, regardless of tectonic setting and the pressure–temperature–time path of the protolith¹⁵.

Unlike the case in the mantle, where gravity-driven compaction can segregate basaltic melt over geologically reasonable timescales^{16,17}, the higher viscosities of crustal melts limit compaction to length scales comparable with the mean grain size of the surrounding matrix^{17,18}. Accordingly, most field evidence points to deformation as the dominant mechanism that segregates and focuses melt flow in the lower crust^{8,9}, and possibly the upper mantle as well (for example, ref. 19). Rock deformation experiments on partially melted granite indicate that at melt fractions (porosities) of 10–40%, pore pressures converge on the mean stress or confining pressure, resulting in macroscopic deformation due to melt-enhanced embrittlement by cataclastic flow^{8,20,21}. These experiments have challenged the long-standing notion of a fundamental rheological threshold (critical melt fraction) below which melt cannot be extracted (for example, ref. 22), and imply that deformation-enhanced segregation can in principle occur at any stage during partial melting^{8,20}.

Mechanically derived melt segregation models have several important implications for geochemical models of granitic magmatism. First, as deformation-assisted melt segregation is efficient in moving melt in the range $10^3 < \eta < 10^6$ from source to local sites of dilation on timescales of approximately 10^{-1} – 10^4 years^{20,23}, melts may not attain chemical or isotopic equilibrium with their surrounding protolith before final extraction and ascent^{23,24}. Second, the proven efficiency of melt segregation in a deformation field makes it unlikely that large, granitic magma chambers will form in the region of partial melting^{12,18}.

Ascent. Gravity is still the most viable driving force for large-scale vertical transport of melt in the continental crust. However, the

traditional idea of buoyant granitic magma ascending through the continental crust as slow-rising, hot Stokes diapirs or by stoping has been largely replaced by models involving the ascent of granitic magmas in narrow conduits, either as self-propagating dykes^{10,25}, along pre-existing faults²⁶ or as an interconnected network of active shear zones and dilational structures^{27,28}. Although the case is still being made for diapiric ascent in the ductile lower crust²⁹, the advantage of dyke/conduit ascent models is that they overcome the severe thermal and mechanical problems associated with transporting very large volumes of magma through the upper brittle continental crust³⁰, as well as explaining the persistence of near-surface plutonism and associated silicic volcanism. However, it remains an open question as to whether plutons are fed by predominantly a few large conduits analogous to basaltic dykes found in ancient shield areas on Earth, dyke swarms similar to those associated with sea-floor spreading, or the pervasive flow of channelled magma^{31,32}.

One of the most striking aspects of the ascent of granitic melt in dykes compared to diapiric rise is the extreme difference in magma ascent rate between both processes, with the former up to a factor of 10^6 faster depending upon the viscosity of the material and conduit width^{25,26}. The narrow dyke widths (~ 1 – 50 m) and rapid ascent velocities predicted by fluid dynamical models have found support in field and experimental studies^{33,34}, while rheological models based on granular flow theory are helping to refine further the ascent rates of magmatic suspensions³⁵. In contrast to diapiric ascent, chemical and thermal interaction between dyke magmas and surrounding country rock will be minimal, and there may be little evidence in the way of geological, geophysical or geochemical evidence to mark the passage of large volumes of granite magma through the crust^{10,25}. Dyke ascent models also bring plutonic granite magmatism more in line with timescales characteristic of silicic volcanism and flood basalt magmatism. The latter provide an interesting corollary, where rapid emplacement at the surface is controlled by flow along deeply penetrating dykes that underlie plateau basalt provinces. These dykes are rarely exposed, but undoubtedly exist.

Emplacement

The emplacement of granitic magma within the continental crust marks the final stage in the granite forming process. Emplacement is defined here as the switch from upward to horizontal flow, and is controlled by a combination of mechanical interactions (either

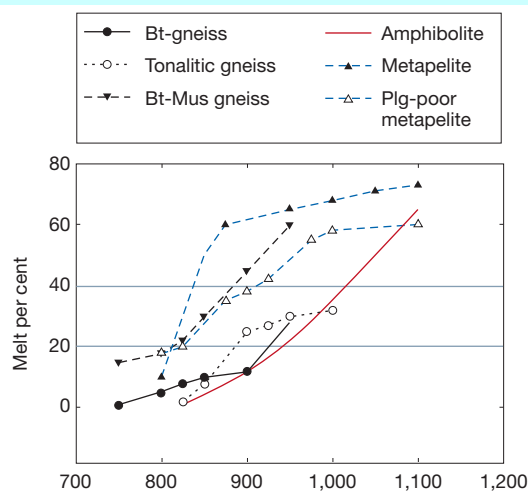


Figure 1 Change in melt fraction (vol. %) as a function of temperature for a range of common crustal rock types undergoing fluid-absent melting (from ref. 9). Note the nonlinear increase in melt fraction with increasing temperature for all rock types. The amphibolite melting curve is a parametrization of seven experimental studies listed in refs 8, 18 and 25. The boxed area shows the range of melt fractions ($F = 0.2$ to 0.4) required to produce most granite compositions. Bt-mus, biotite-muscovite; plg poor, plagioclase-poor; metapelite, metamorphosed pelite.

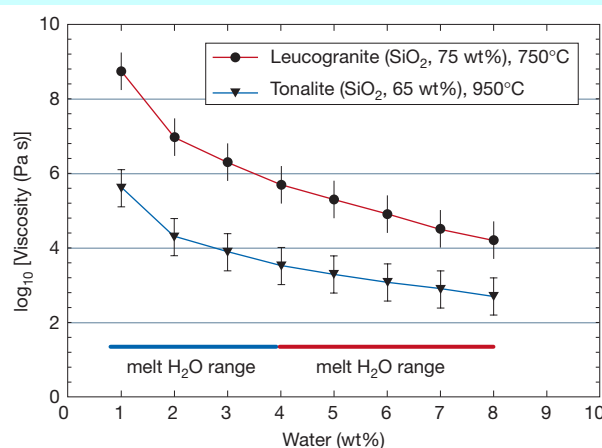
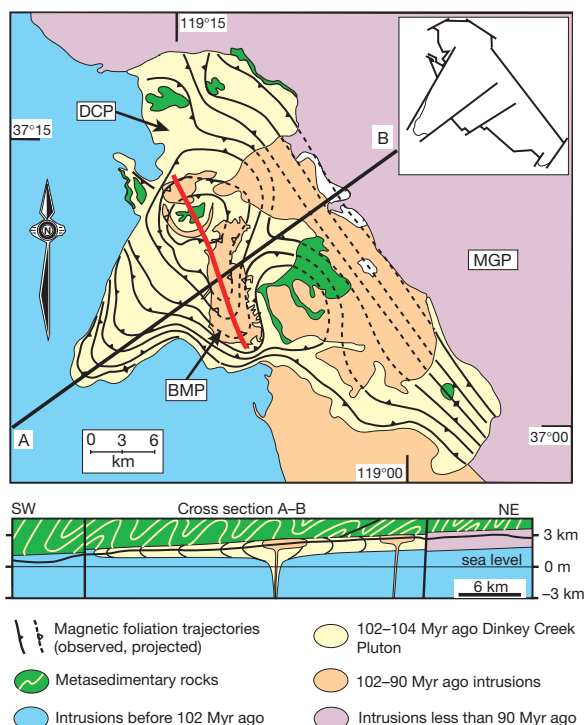


Figure 2 Melt viscosity as a function of melt water (wt%) content for typical tonalite and leucogranite liquid compositions (data from ref. 15 and references therein) at a fixed pressure of 800 MPa. Blue and red horizontal lines show the range of water contents typical for natural melts. The estimated $\log_{10}$ values of the median viscosities (in Pa s) of the liquids at their ‘ideal’ water contents of 4 wt% (tonalite) and 6 wt% (leucogranite) are 3.8 and 4.9 respectively.

pre-existing or emplacement-generated wall rock structures), and density effects between the spreading flow and its surroundings^{36,37}. The way in which rocks make way for this newly incoming magma (the so-called space problem)¹⁴ has challenged geologists for most of the twentieth century. The problem becomes particularly acute where batholithic volumes ($\geq 1 \times 10^5 \text{ km}^3$) of magma are considered to have been emplaced in a single episode, often exemplified in models of diapiric rise, where the processes of ascent and emplacement are blurred²⁵. New ideas that have helped improve this are the recognition of the important role played by tectonic activity in making space in the crust for incoming magmas (see ref. 37), more realistic interpretations of the geometry of granitic intrusions at depth (for example Box 1), and the recognition that

emplacement is an episodic processes involving discrete pulses of magma. Structural, analogue and numerical models^{38–41} indicate that space for incoming magma at strain rates of between 10^{-15} (average lithospheric strain rate) and 10^{-10} s^{-1} (several orders of magnitude greater than the average lithospheric strain rate) can be achieved through a combination of lateral fault opening, roof lifting and lowering of the growing magma chamber floor. In the latter two mechanisms, space is created ultimately by surface uplift and erosion or by volume reduction caused by melt extraction in the source⁴⁰, possibly aided by isostatic depression of the Moho⁴². Such models resolve the insufficient strain record in the vicinity of certain plutons and the slow tectonic widening rates on bounding faults that have led some to question the ability of the crust to make space for incoming magma by fault dilation⁴³.

Box 1



Geological map and cross section of the 800 km², 102–104-Myr-old Dinkey Creek pluton, Sierra Nevada batholith, California. The pre-erosion thickness of the lobe structure in the southwest of the pluton is estimated to be between 900 and 3,700 m, based on fluid mechanical analysis of the magnetic fabric pattern, which also indicates a feeder zone trending to the north-northwest (red line on map)^{50,51}. Evidence for a relatively flat roof above the pluton is provided by remnants of metasedimentary rocks preserved at high elevations. Structure contours on the contacts of these roof pendants indicate that the current roof dips about 10° SW. This is similar to the Cenozoic tilt of the adjacent Mount Givens pluton (MGP) determined by palaeomagnetic studies, suggesting that the Dinkey Creek pluton (DCP) roof was originally horizontal and located, on average, about 400 m above the present erosion surface. Geobarometry indicates that the pluton was emplaced at a palaeodepth of about 15 km (ref. 50). Repeated linear contact orientations, displacements of roof structure contours, and geometry of roof pendant contacts, suggest that the majority of the DCP's vertical contacts, and irregularities in its roof, were controlled by pre-existing NW-, NE-, N- and E-trending faults and fractures, which acted as guides for vertical displacement of the pluton floor (see inset). Similarly oriented structures controlled the shape of the 40 km² Bald Mountain pluton (BMP) and the 1,500 km² MGP.

Three-dimensional shapes of granitic intrusions

Important information on the emplacement mechanism of granitic magmas is preserved in the three-dimensional (3D) shape of the crystallized pluton. Detailed geophysical (gravity and seismic) data now allow the subsurface 3D geometries of granitic plutons to be imaged with a high degree of confidence. The majority of plutons so far investigated appear as flat-lying to open funnel-shaped structures with central or marginal feeder zones, consistent with an increasing number of field studies that find plutons to be internally sheeted on the decimetre to kilometre scale⁴⁴. Gravity models based on detailed measurements made at the same spatial resolution as geological observations are now common. Depth determination, after gravity data inversion, structural measurements from field observations and determination of petrofabrics using the anisotropy of magnetic susceptibility (AMS) technique⁴⁵, can be combined with geochemical information on magma evolution to build up a comprehensive 3D picture of pluton geometry⁴⁶. Such multidisciplinary investigations again show that the common form is sheet-like as do the few detailed seismic surveys so far undertaken over or through granitic batholiths⁴⁷.

Mechanism of pluton growth

Empirical studies of pluton dimensions, based on field and geophysical measurements, suggest that the growth of a laterally

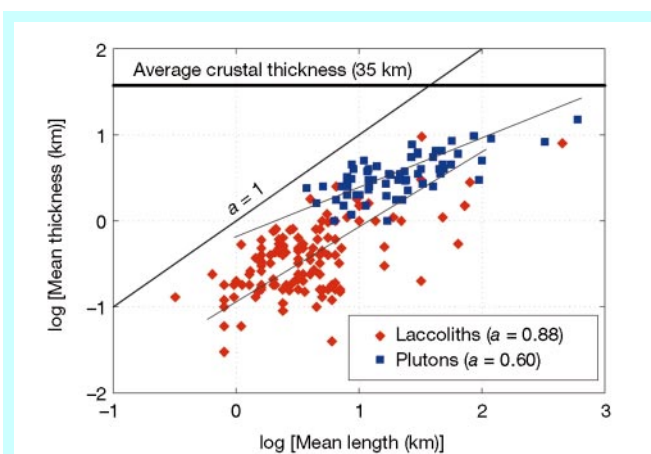


Figure 3 Mean (vertical) thickness versus mean (horizontal) length for granitic plutons and laccoliths. Reduced major-axis regression defines a power-law curve for plutons with an exponent a of 0.6 ± 0.1 . Laccoliths (shallow-level intrusions) are described by a power-law exponent of 0.88 ± 0.1 (ref. 48). The line $a = 1$ defines the critical divide between predominantly vertical inflation ($a > 1$) and predominantly horizontal elongation ($a < 1$) during intrusion growth. Significantly different power-law exponents rule out a simple genetic relationship between both populations. Differences may be due to mechanical effects, with limits in thickness reflecting floor depression (plutons) and roof lifting (laccoliths).

spreading and vertically thickening intrusive flow evolves according to a power-law (self-affine) relationship of the form $L = kT^a$, typical of systems exhibiting scale-invariant (fractal) behaviour⁴⁸. The inherent preference for scale-invariant tabular sheet geometries in granitic plutons from a variety of tectonic settings is shown in Fig. 3. This relationship is best explained in mechanical terms by the need for magma at the emplacement level to travel horizontally some distance before vertical thickening can occur, either by hydraulic lifting of its overburden (in the case of shallow-level intrusions, known as laccoliths) or sagging of its floor. Plutons thus undergo a birth stage, characterized by lateral spreading, followed by an inflation stage marked by vertical thickening. Although from Fig. 3, plutons and laccoliths do not form a simple continuum as previously thought⁴⁸, the tabular sheet model suggests that larger plutons grow from smaller ones according to a power-law inflation growth curve, ultimately to form crustal-scale batholithic intrusions^{40,48}. Evidence of this growth process is recorded in the 1,200-km-long Coastal batholith of Peru, one of the best-studied arc batholiths on Earth. Here, combined field, petrological,

geochemical and geophysical (gravity) studies show that on a crustal scale the exposed batholith forms a thin (3–7-km-thick) low-density layer that coalesced from numerous smaller plutons with aspect ratios of between 17–20:1 (ref. 49). Detailed studies of the Sierra Nevada batholith of California reveal a similar picture, in which batholith construction occurred by intrusion of 2 to 2,000 km² granitic plutons between 120 and 80 Myr ago⁵⁰. The Dinkey Creek pluton is typical of the intrusions that make up the central Sierra Nevada batholith (Box 1) and many other continental magmatic arcs. Field mapping of contacts, roof pendants and internal fabrics, combined with AMS studies, indicate that the pluton was emplaced as a 900–3,700-m-thick tabular intrusion with vertical walls and a gently dipping roof⁵¹. Magma was fed into the pluton by a conduit trending north-northwest, with space created by vertical inflation and depression of the floor caused by translation on pre-existing fractures in the wall rocks.

Timescales of pluton growth

The emerging picture of plutons, with tabular 3D geometry and growth by vertical displacements of their roofs and floors, allows us to place limits on the rates and times of their emplacement (Fig. 4). If we assume that a disk-shaped pluton grows according to the empirical power-law relation shown in Fig. 3

$$T = 0.6(\pm 0.15)L^{0.6\pm 0.1} \tag{1}$$

then its filling time can be estimated when the volumetric filling rate, Q , is known. Taking conservative values of magma viscosities (Fig. 2), wall-rock/magma density differences and feeder dyke morphologies gives a range of Q from 0.01 to 100 m³ s^{−1}, and places lower and upper bounds on pluton filling times of less than 40 days to more than 1 Myr for plutons less than 1 km to more than 100 km across (Fig. 4). The thicknesses of the plutons highlighted in Box 1 have been estimated using the above power-law relationship, given the horizontal area and equivalent L value for each pluton. Taking the median Q value of 1 m³ s^{−1} gives ranges of net emplacement times of about 10³, 10⁴ and 10⁵ yr for the Bald Mountain, Dinkey Creek and Mount Givens plutons, respectively (Fig. 4). At the fastest magma delivery rates, all three plutons could have been emplaced in less than 1,000 yr. Similar timescales have been proposed for Himalayan granites⁵². We also note that the pluton thicknesses may have been over estimated by equation (1), as indicated by comparison to independent calculations of the thickness of the Dinkey Creek pluton lobe⁵¹, resulting in faster emplacement times for all three cases (Fig. 4).

Towards a unified model for granite magmatism

The formation of granite intrusions in the middle to upper crust is

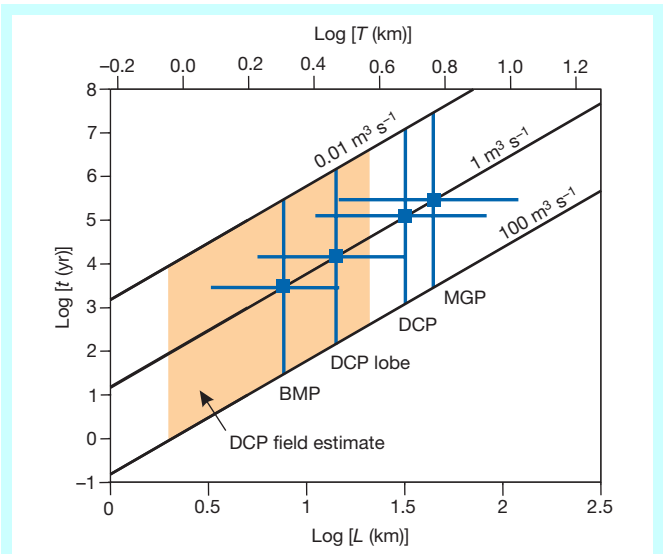


Figure 4 Estimated filling times for tabular disk-shaped plutons. Their thickness (T) to width (L) ratio is given by equation (1) for a range of permissible filling rates (Q , m³ s^{−1}). Heavy horizontal lines are the thickness ranges of plutons in Box 1, estimated using equation (1). Vertical lines are the ranges of their possible filling times, bracketed by the filling rates. The coloured prism indicates the range of thicknesses estimated independently for the southwest lobe of the DCP using structural data (Box 1).

Table 1 The four processes responsible for granitic magmatism in the continental crust, and estimated timescales				
Process	Tectonic setting	Mechanism	Timescale (years)	References
Partial melting				
Fluid present (high a_{H_2O})	Transpressional/ transtensional orogens	Crustal thickening and decompression/ asthenospheric upwelling	$>10^5$	4, 5, 52
Fluid absent	Magmatic arcs (extensional/compressional)	Magmatic under/intraplating	$10^2\text{--}10^5$	4, 6, 7, 12
Segregation				
	Extensional/compressional	Gravity-driven compaction	$10^5\text{--}10^9$	16–18
		Deformation-enhanced flow/fracturing	$10^6\text{--}10^3$	16–18, 20, 21
Ascent				
Dyke/conduit flow	Mainly extensional	Buoyancy or deformation-assisted flow	$10^{-1}\text{--}10^2$	10, 25, 26, 34
Pervasive flow	Transpressional	Viscous flow in hot permeable crust	$\leq 10^6$	27, 28, 31, 32
Diapiric rise	Extensional/compressional	Buoyancy	$10^5\text{--}10^9$	29, 30, 43
Emplacement				
	Extensional/compressional	Entrainment along structural/ rheological traps, buoyancy	$10^2\text{--}10^4$	33, 36–40, 51, 52

governed by four discrete processes, summarized in Table 1. We have a number of reasons for confidence that these processes, running in series, provide an internally consistent qualitative model for granite magmatism in the Earth. First, field observations made in a variety of tectonic settings on rocks of different ages show that structural features relating to magma segregation, emplacement and pluton shape are repeated in space and time. We also find it encouraging that the high magma flow rates predicted in dyke ascent models can be reconciled with mechanical models for emplacement of tabular plutons at geologically realistic strain rates. Finally, tabular to funnel-shaped pluton geometries with inclined feeder zones predicted by fluid dynamical and structural models are consistent with recent high-resolution gravity and seismic investigations⁴⁶. The rate-limiting step in granite magmatism is the timescale of partial melting^{3,52}; the follow-on stages of segregation, ascent and emplacement can be geologically extremely rapid—perhaps even catastrophic.

Future experimental and numerical modelling of the underlying physics of each stage, and their associated feedback relations, will help to constrain further the rates at which these processes operate, and lead ultimately to models that capture the full complexity of granitic magma systems on the crustal scale. Problems yet to be tackled are numerous, and include matching predicted melting rates based on thermal models with kinetic studies of mineral reactions, and identification of a satisfactory focusing mechanism that enables partial melt distributed over a wide source region to be drained by a small number of conduits. □

1. Taylor, S. R. & McLennan, S. M. The geochemical evolution of the continental crust. *Rev. Geophys.* **33**, 241–265 (1995).
2. Brown, M. The generation, segregation, ascent and emplacement of granite magma: the migmatite-to-crustally-derived granite connection in thickened orogens. *Earth Sci. Rev.* **36**, 83–130 (1994).
3. Petford, N., Clemens, J. D. & Vigneresse, J. L. in *Granite: From Segregation of Melt to Emplacement* (eds Bouchez, J. L., Hutton, D. H. W. & Stephens, W. E.) 3–10 (Kluwer, Dordrecht, 1997).
4. Thompson, A. B. in *Understanding Granites: Integrating New and Classical Techniques* (eds Castro, A., Fernandez, C., Vigneresse, J. L.) 7–25 (Geological Society of London Special Publication 158, 1999).
5. Thompson, A. B. & Connolly, J. A. D. Melting of the continental crust: some thermal and petrological constraints on anatexis in continental collision zones and other tectonic settings. *J. Geophys. Res.* **100**, 15556–15579 (1995).
6. Huppert, H. E. & Sparks, R. S. J. The generation of granitic magmas by intrusion of basalt into continental crust. *J. Petrol.* **29**, 599–642 (1988).
7. Bergantz, G. W. Underplating and partial melting: implications for melt generation and extraction. *Science* **254**, 1039–1095 (1989).
8. Brown, M. & Rushmer, T. in *Deformation-Enhanced Fluid Transport in the Earth's Crust and Mantle* (ed. Holness, M.) 111–144 (Chapman & Hall, London, 1997).
9. Vigneresse, J. L., Barbey, P. & Cuney, M. Rheological transitions during partial melting and crystallisation with application to felsic magma segregation and transfer. *J. Petrol.* **37**, 1579–1600 (1996).
10. Clemens, J. D. & Mawer, C. K. Granitic magma transport by fracture propagation. *Tectonophysics* **204**, 339–360 (1992).
11. Watt, G. R., Oliver, N. H. S. & Griffin, B. J. Evidence for reaction-induced microfracturing in granulite facies migmatites. *Geology* **28**, 327–330 (2000).
12. Petford, N. & Koenders, M. A. Self-organisation and fracture connectivity in rapidly heated continental crust. *J. Struct. Geol.* **20**, 1425–1434 (1998).
13. Dingwell, D. B., Bagdassarov, N. S., Bussod, G. Y. & Webb, S. L. in *Experiments at High Pressure and Applications to the Earth's Mantle* (ed. Luth, R. W.) 131–196 (Short course Vol. 21, Mineralogical Association of Canada, 1993).
14. Pitcher, W. S. *The Nature and Origin of Granite* (Blackie Academic, Glasgow, 1993).
15. Clemens, J. D. & Petford, N. Granitic melt viscosity and silicic magma dynamics in contrasting tectonic settings. *J. Geol. Soc. Lond.* **156**, 1057–1060 (1999).
16. McKenzie, D. The generation and compaction of partially molten rocks. *J. Petrol.* **25**, 713–765 (1984).
17. Wickham, S. M. The segregation and emplacement of granitic magmas. *J. Geol. Soc. Lond.* **144**, 281–297 (1987).
18. Petford, N. Segregation of tonalitic-trondhjemitic melts in the continental crust: the mantle connection. *J. Geophys. Res.* **100**, 15735–15743 (1995).
19. Kelemen, P. B. & Dick, H. J. B. Focused flow and localized deformation in the upper mantle: juxtaposition of replacive dunite and ductile shear zones in the Josephine peridotite, SW Oregon. *J. Geophys. Res.* **100**, 423–438 (1995).

20. Rutter, E. H. & Neumann, D. H. K. Experimental deformation of partially molten Westerly granite under fluid-absent conditions with implications for the extraction of granitic magmas. *J. Geophys. Res.* **100**, 15697–15715 (1995).
21. Davidson, C., Schmid, S. M. & Hollister, L. S. Role of melt during deformation in the deep crust. *Terra Nova* **6**, 133–142 (1994).
22. Arzi, A. A. Critical phenomena in the rheology of partially molten rocks. *Tectonophysics* **44**, 173–184 (1979).
23. Sawyer, E. W. Disequilibrium melting and the rate of melt-residuum separation during migmatization of mafic rocks from the Grenville front, Quebec. *J. Petrol.* **32**, 701–38 (1991).
24. Davies, G. R. & Tommasini, S. Isotopic disequilibrium during rapid crustal anatexis: implications for petrogenetic studies of magmatic processes. *Chem. Geol.* **162**, 169–191 (2000).
25. Clemens, J. D., Petford, N. & Mawer, C. K. in *Deformation-Enhanced Fluid Transport in the Earth's Crust and Mantle* (ed. Holness, M.) 145–172 (Chapman & Hall, Lond. 1997).
26. Petford, N., Kerr, R. C. & Lister, J. R. Dike transport of granitoid magmas. *Geology* **21**, 845–848 (1993).
27. D'Lemos, R. S., Brown, M. & Strachan, R. A. Granite magma generation, ascent and emplacement within a transpressional orogen. *J. Geol. Soc. Lond.* **149**, 487–490 (1993).
28. Collins, W. J. & Sawyer, E. W. Pervasive granitoid magma transport through the lower-middle crust during non-coaxial compressional deformation. *J. Metamorph. Geol.* **14**, 565–579 (1996).
29. Weinberg, R. F. & Podladchikov, Y. Y. Diapiric ascent of magmas through power law crust and mantle. *J. Geophys. Res.* **99**, 9543–9559 (1994).
30. Marsh, B. D. On the mechanics of igneous diapirism, stoping and zone melting. *Am. J. Sci.* **282**, 808–855 (1982).
31. Brown, M. & Solar, G. S. The mechanism of ascent and emplacement of granite magma during transpression: a syntectonic granite paradigm. *Tectonophysics* **312**, 1–33 (1999).
32. Weinberg, R. F. Mesoscale pervasive felsic magma migration: alternatives to dyking. *Lithos* **46**, 393–410 (1999).
33. Scaillet, B., Pecher, A., Rochette, P. & Champenois, M. The Gangotri granite (Garhwal Himalaya): laccolith emplacement in an extending collisional belt. *J. Geophys. Res.* **100**, 585–607 (1994).
34. Brandon, A. D., Chacko, T. & Creaser, R. A. Constraints on rates of granitic magma transport from epidote dissolution kinetics. *Science* **271**, 1845–1848 (1996).
35. Petford, N. & Koenders, M. A. Granular flow and viscous fluctuations in low Bagnold number granitic magmas. *J. Geol. Soc. Lond.* **155**, 873–881 (1998).
36. Hogan, J. P. & Gilbert, M. C. The A-type Mount Scott granite sheet: importance of crustal magma traps. *J. Geophys. Res.* **100**, 15799–15792 (1995).
37. Hutton, D. H. W. Granite emplacement mechanisms and tectonic controls: inferences from deformation studies. *Trans. R. Soc. Edinb.: Earth Sci.* **79**, 245–255 (1988).
38. Fernandez, C. & Castro, A. Pluton accommodation at high strain rates in the upper continental crust. The example of the Central Extremadura batholith, Spain. *J. Struct. Geol.* **21**, 1143–1149 (1999).
39. Benn, K., Odonne, F. & de Saint Blanquat, M. Pluton emplacement during transpression in brittle crust: new views from analogue experiments. *Geology* **26**, 1079–1082 (1998).
40. Cruden, A. R. On the emplacement of tabular granites. *J. Geol. Soc. Lond.* **155**, 853–862 (1998).
41. Roman-Berdiel, T., Gapais, D. & Brun, J. P. Granite intrusion along strike-slip zones in experiment and nature. *Am. J. Sci.* **297**, 651–678 (1997).
42. Brown, E. H. & McClelland, W. C. Pluton emplacement by sheeting and vertical ballooning in part of the southeast Coast Plutonic Complex, British Columbia. *Geol. Soc. Am. Bull.* **112**, 708–719 (2000).
43. Miller, R. B. & Paterson, S. R. In defence of magmatic diapirs. *J. Struct. Geol.* **21**, 1161–1173 (1999).
44. Grocott, J., Garde, A., Chadwick, B., Cruden, A. R. & Swager, C. Emplacement of Rapakivi granite and syenite by floor depression and roof uplift in the Paleoproterozoic Ketildian orogen, South Greenland. *J. Geol. Soc. Lond.* **156**, 15–24 (1999).
45. Bouchez, J. L., Hutton, D. H. W. & Stephens, W. E. (eds) *Granite: From Segregation of Melt to Emplacement* (Kluwer, Dordrecht, 1997).
46. Améglio, L. & Vigneresse, J. L. in *Understanding Granites: Integrating New and Classical Techniques* (eds Castro, A., Fernandez, C., Vigneresse, J. L.) 39–54 (Geological Society of London Special Publication 158, 1999).
47. Evans, D. J. et al. Seismic reflection data and the internal structure of the Lake District batholith, Cumbria, northern England. *Proc. Yorkshire Geol. Soc.* **50**, 11–24 (1994).
48. McCaffrey, K. J. W. & Petford, N. Are granitic intrusions scale invariant? *J. Geol. Soc. Lond.* **154**, 1–4 (1997).
49. Atherton, M. P. Shape and intrusion style of the Coastal batholith, Peru. In *4th Int. Symp. on Andean Geodynamics* (extended abstr.) 60–63 (1999).
50. Bateman, P. C. Plutonism in the central part of the Sierra Nevada batholith, California. *US Geol. Surv. Prof. Pap.* **1483** (1992).
51. Cruden, A. R., Tobisch, O. T. & Launeau, P. Magnetic fabric evidence for conduit-fed emplacement of a tabular intrusion: Dinkley Creek pluton, central Sierra Nevada batholith, California. *J. Geophys. Res.* **104**, 10511–10530 (1999).
52. Harris, N., Vance, D. & Ayres, M. From sediment to granite: timescales of anatexis in the upper crust. *Chem. Geol.* **162**, 155–167 (2000).

Acknowledgements

We would like to thank M. Brown and A. Brandon for helpful reviews and J. Clemens and M. Atherton for useful discussions during the preparation of this manuscript.

Correspondence and requests for materials should be addressed to N.P. (e-mail: n.pet@kingston.ac.uk).

Fossiliferous Lana'i deposits formed by multiple events rather than a single giant tsunami

Ken H. Rubin, Charles H. Fletcher III & Clark Sherman

Department of Geology and Geophysics, SOEST, University of Hawaii, 2525 Correa Road, Honolulu, Hawaii 96822, USA

Giant tsunamis, generated by submarine landslides in the Hawaiian Islands, have been thought to be responsible for the deposition of chaotic gravels high on the southern coastal slopes of the islands of Lana'i and Moloka'i, Hawaii. Here we investigate this hypothesis, using uranium–thorium dating of the Hulopoe gravel (on Lana'i) and a study of stratigraphic relationships, such as facies changes and hiatuses, within the deposit. The Hulopoe gravel contains corals of two age groups, representing marine isotope stages 5e and 7 (~135,000 and 240,000 years ago, respectively), with significant geographical and stratigraphic ordering. We show that the Hulopoe gravel was formed by multiple depositional events, separated by considerable periods of time, thus invalidating the main premise of the 'giant wave' hypothesis. Instead, the gravels were probably deposited during interglacial periods (when sea level was relatively high) by typical Hawaiian shoreline processes such as seasonal wave patterns, storm events and possibly 'normal' tsunamis, and reached their present height by uplift of Lana'i.

Lana'i is an eroded, extinct shield volcano about 1 Myr old, overlain at lower elevations by discontinuous sedimentary rock outcrops. Early detailed studies reported bioclastic gravels on the island's southern slopes to 60–70 m elevation (primarily preserved in gullies and canyons of the arid coastal region; Fig. 1) and much less extensive, isolated outcrops up to ~170 m elevation¹. The largest concentration of coral-bearing conglomerates is a 0.6 km² area upslope of Hulopoe and Kapihaa Bays¹ (Fig. 1); the thickest deposits (~10 m) are near the present shoreline. A site within Kapihaa gully was later named the "Hulopoe gravel" type locality². These deposits were originally attributed to a succession of relative glacioeustatic marine high stands³. Given the elevation range of the deposits, uplift of Lana'i would be required. Besides the main deposits, two small higher-elevation, presumed older, fossiliferous limestone localities occur. These are isolated conglomerates at 190 m in Kaluakapo crater and vein-filling calcite <1 cm wide by <1 m long at 326 m; both localities rest astride upthrown (relative) blocks of the normal-fault-bounded fossil south rift zone of the Lana'i shield¹. This, and a change in soil type and thickness and rounded boulders at 326–375 m were also interpreted as a relative marine high stand, with unspecified uplift mechanism^{1,3}.

An alternative, dominantly accepted hypothesis^{2,4} suggests that a "giant wave" tsunami² (or tsunami wave train⁴) formed the Hulopoe gravel. Wave run-up to ~375 m (refs 2 and 4) was to have stripped soil and arranged basalt boulders on drainage interfluvies. This single, catastrophic, 'above sea level' event was proposed because tidal gauge data and ages of drowned fossil reefs indicated subsidence on the southernmost islands of Hawaii (2.4 mm yr⁻¹) and southern Maui (0.3 mm yr⁻¹)^{5,6}, thus suggesting that now-emergent palaeo-shoreline deposits were not possible on Lana'i^{2,4}. However, other than the circumstances of the proposed tsunami there is no published evidence (such as stratigraphic or geochronological evidence) that links the deposits comprising the Hulopoe gravel together or to other geological features on Lana'i. The event was 'dated' at 105 kyr ago using coral clast U-series ages determined by traditional α -counting methods (one at 134 kyr ago from 155 m in Kaluakapo crater and one each at 108 and 101 kyr ago from 115–120 m in Kawaiu gulch). None of the dated clasts were from the type locality, nor were stratigraphic correlations to the type locality presented^{2,4}. A giant wave is a plausible mechanism for forming

chaotic deposits like the Hulopoe gravel^{7,8} and has been invoked for other features globally (for example, the 105 kyr ago 'Lana'i' wave was proposed to have caused widespread coastal erosion in western Australia)⁹. A bioclastic deposit exposed to 65 m elevation on Moloka'i has also been attributed to a giant tsunami (at 200–240 kyr ago)¹⁰.

This project is a test of Hulopoe gravel depositional scenarios using mapping and radiometric ages of clasts. The giant wave scenario should not result in a meaningful relationship between clast ages and distribution through the Hulopoe gravel, as source material would be randomly mobilized and deposited. The successive marine high stands scenario should result in an upslope age progression as long as uplift outpaced sea level change¹¹. Downslope reworking of clasts from older deposits and depositional succession would also cause down-section age increases at lower elevations.

Hulopoe gravel outcrops and stratigraphy

To better define the origin(s) of the Hulopoe gravel we began in 1995 to map, sample and date clasts from Kapihaa gully, an unnamed gully immediately to the west (referred to here as Archeology gully) and the coastal headlands extending to Kaluakoi point (Fig. 1). Our earliest stratigraphy and five high-resolution thermal ionization mass spectrometry (TIMS) ²³⁰Th–²³⁴U–²³⁸U analyses confirmed a Pleistocene age for the coral clasts (marine isotope stages 5e and 7), indicated more than three mappable units with distinct sedimentary facies, and suggested several depositional events¹². We have since completed additional mapping, sampling and dating; in addition, a sedimentological study of the type locality section was carried out¹³. Conglomerates in and around the type locality include lenticular beds of a range of cementation characteristics (dense micrite to mud), and a wide variety of clast types, sizes and shapes. None of the mapped deposits appear to be *in situ* reef. Instead, clast size and matrix variations indicate original and reworked littoral and alluvial deposits from a spectrum of medium to high-energy environments.

Our mapping was accomplished using EDM (electronic distance meter) surveys, with results similar to those of Stearns¹. Our results differ significantly from later reports that show a continuous apron extending from sea level to ~100 m elevation². Our observations differ in two other specific ways that bear on their origin: (1) we do find extensive deposition on coastal headlands near Kapihaa gully;

and (2) in three gullies examined we do not find deposits concentrated on the west sides⁴ (for example, most Archeology gully gravels are on the eastern wall). Supposed deposit thickening on western sides of gullies and lack of deposition on headlands was argued to indicate an oblique giant wave deposition angle and backwash stripping of headlands, respectively⁴. Our multiple unit stratigraphy¹² also differs from the tsunami model^{2,4} (which has one stratigraphic unit composed of six beds of alternating coarseness and coral abundance, proposed to result from run-up and backwash of a three-wave giant tsunami). Eight stratigraphic units have since been identified at the type locality, including three subaerial time break discontinuities (with truncated palaeosols and root casts)¹³. Our geochronological system of sampling is consistent with the

newest stratigraphy, and includes a thick sequence of poorly-sorted conglomerates organized in the lower elevations of Archeology and Kapihaa gullies into two reversely graded sequences dominated by subrounded to subangular clasts of basalt ($\geq 80\%$) and coral (0–20 vol.%). Coral clasts are conspicuously absent from these beds in two 75–100 m stretches along Kapihaa gully, although other apparently non-marine calcareous clasts (for example, pieces of caliche) occur, indicating discontinuity of bioclast deposition. These beds overlie a well-cemented bioclastic basal unit in Kapihaa gully that at 18–25 m elevation forms prominent gully floor outcrops. In other places, bedrock basalt floors the gully. An important feature of the type locality stratigraphy is a red-brown mud/sand matrix-supported basalt gravel layer just above our basal unit (units A and B in

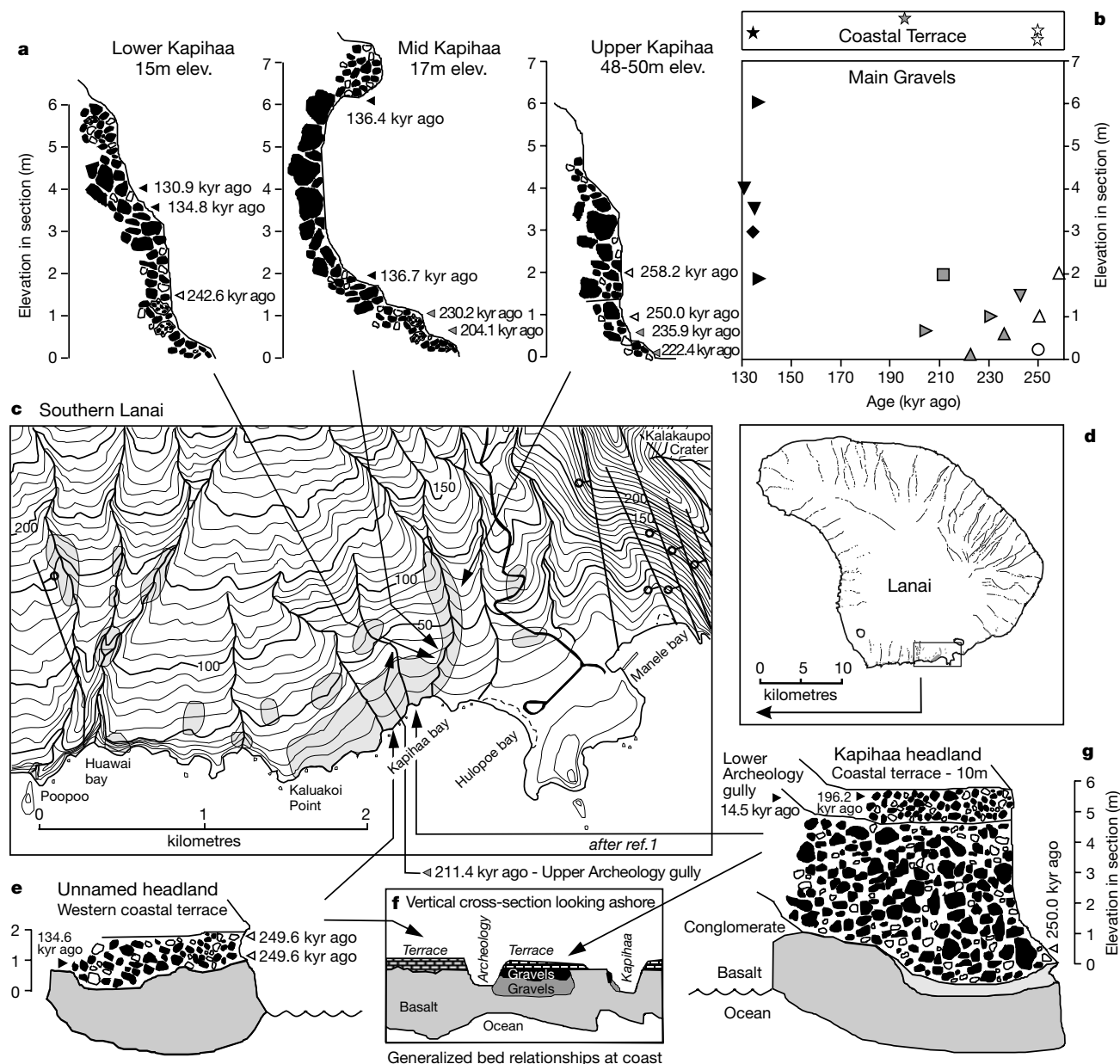


Figure 1 Summary of the study area on the southern coast of Lana'i, Hawaii. **a, e, g**, Graphical representations of each outcrop studied (basalt clasts are black, corals are white, and basalt bedrock on the headlands is grey). Sampling elevations within the sections are depicted by arrowheads with Th–U ages as labels. **c**, Map with sampling sites for coral clast dating (10-m contour interval). Outcrops of Hulopoe gravel are shown in light grey, normal faults are depicted by lines with a circle on the down-dropped side¹. **d**, Map of Lana'i. The area depicted in **c** is enclosed within the rectangle. **f**, Representation

of the general relationships between gravel beds (black, dark grey), basalt basement (light grey), and the coastal terrace (white bricks). **b**, Composite plot of age versus elevation in section. Symbols in this panel are keyed to outcrop location: lower (inverse triangles), mid (right triangles), and upper (triangles) Kapihaa gully; lower (diamonds) and upper (squares) Archeology gully; Kapihaa headland (circles); coastal terrace (stars). **a–g**, Black and grey filled symbols represent most reliable stage 5e and stage 7 ages (respectively) and open symbols represent least reliable ages (see Methods for tests of age reliability).

ref. 13), interpreted as an alluvial deposit and bounded by two of three “time-gap” disconformities of the sequence¹³.

This entire sequence is overlain in the near-coastal region by what we term a ‘Coastal Terrace’ deposit, which is a weakly lithified unit with calcareous cement containing a somewhat higher percentage of coral clasts (30–40%), as well as mollusc and gastropod shells, foraminifera, and lithified beach rock (Fig. 1e–g). Importantly, this lithologically distinct unit clearly overlies all other gravel units near the mouths of gullies, mantles bedrock basalt unconformably at nearby coastal headlands and sea cliffs¹², and contains clasts that were probably reworked from stratigraphically lower beds¹³. The terrace mantles the other deposits and basalts uniformly and was thus clearly deposited in a separate event with significantly different preservation characteristics and a different immediate origin from that of the other gravel beds. Where the terrace overlies gravels at the coast it is uniformly 1 m thick but locally reaches 2 m thick over bedrock, infilling pre-existing topography. It has a uniform 6–8° seaward dip, forming a relatively constant elevation terrace at equal distance back from the sea, and on headlands is commonly undercut by erosion at its base, forming a visor. It is restricted to near the coast and in Archeology gully can be traced as a continuous layer inland to ~100 m from Kapihaa headland (Fig. 1).

Thus, the previously named Hulopoe gravel is a collection of several units that we group into ‘Basal’, ‘Middle’ and ‘Coastal Terrace’ for ease of discussion. Together the basal and middle packages form the ‘Main body’ of the Hulopoe gravel (MB). We sampled five locales within the MB, a series of coastal terrace sites where it overlies basalt to the west of Archeology gully, and a modern lithified shoreline gravel (LSG) of subrounded basalt clasts, coral clasts, and coralline algae. No suitable samples for dating were found at the Hulopoe gravel type locality outcrop, although our lower and mid Kapihaa sites bracket it by ~75 m on either side. Dated clasts were massive coral forms (primarily *Porites lobata*) excavated from well-lithified strata.

Geochronology

The ages of 18 Hulopoe gravel coral clasts were determined by high-

resolution TIMS ²³⁰Th–²³⁴U–²³⁸U analyses (Table 1). Calculated ages fall into two groupings: 130.9–136.7 kyr ago, and 196.2–258.2 kyr ago (Fig. 2). These correspond to the last two interglacial marine high stands and were times of enhanced coral growth/reef preservation in Hawaiian waters^{14–16}. Three clasts from the LSG were also dated late Holocene (2.0–3.2 kyr ago), correlating with the age of the Kapapa marine high stand¹⁷.

Eight of the 18 Hulopoe gravel samples were analysed in replicate, agreeing to 0.06 to 0.8%. The stage 7 group age range is larger than expected from analytical reproducibility (the best and worst replicates were in this group and replicate to 0.2 to 3% of the overall age range). Analytical errors do not appear to have contributed significantly to the data spread, but geological errors may very well have done so, owing to the complex depositional history and the present conditions of subaerial exposure. Broadly speaking, all of the dated specimens show evidence for meteoric diagenesis (exterior surface discolouration, recrystallization to calcite, Fe-rich mineralization on interior domains, minor dissolution textures) as well as a second phase of either meteoric or marine diagenesis (micritic cementation). Although we took great care to obtain the most-pristine appearing part of each sample, almost certainly some altered material was analysed, and some ages may thus be less reliable than others. Commonly employed geochemical tests of closed-system isotopic evolution since coral death are described in the Methods.

Samples with reliable stage 5e ages (based primarily on calculated initial ²³⁴U/²³⁸U) span 130.9 to 136.7 kyr. Many of these ages overlap with a recently suggested age for the penultimate 5e deglaciation (135 ± 2.5 kyr ago)¹⁸ and with coral ages within the *in situ* 5e reef formation on O’ahu^{14,15}. Some but not all of the stage 7 age range is likely to be real: the most reliable stage 7 samples are in two clusters (196.2 to 211.4 kyr ago, three sample mean is 204 kyr ago and one sample was 230.2 kyr ago). Two other less reliable ages (222.4 and 235.9 kyr ago) are sufficiently close to the latter to suggest that they are also valid (mean of second three-sample cluster is 230 kyr ago). The younger cluster correlates with late in the stage 7 interglacial (events 7.1–7.3) and the older cluster correlates with event 7.5¹⁹.

Table 1 Sample particulars, U-series data, ages and 2σ uncertainties

Location and samples	Site	Elevation (m)	Calcite %	U (ng g ⁻¹)	Th (pg g ⁻¹)	²³² Th/ ²³⁸ U atom ratio × 10 ⁵	(²³⁰ Th/ ²³⁸ U) activity ratio	δ ²³⁴ U measured	Age (kyr ago)	δ ²³⁴ U _i initial
Hulopoe bay										
Lan2-1-2	Tidal	0.50	0	2,612 ± 7	101 ± 0.4	4.01 ± 0.02	0.0210 ± 0.0007	146.1 ± 2.3	2.01 ± 0.07	147 ± 2
Lan2-1-1	Tidal	0.50	0	3,248 ± 8	379 ± 0.9	12.1 ± 0.04	0.0213 ± 0.0006	147.1 ± 1.5	2.04 ± 0.06	148 ± 1
Lan2-1-4	Tidal	0.50	0	2,692 ± 7	198 ± 0.4	7.60 ± 0.02	0.0334 ± 0.0007	145.0 ± 3.2	3.22 ± 0.07	146 ± 3
Coastal										
Lan3-3-1	Western	3.8	<2.3	2,861 ± 7	287 ± 0.5	10.4 ± 0.03	0.795 ± 0.004	102.8 ± 2.2	134.6 ± 0.9	151 ± 3
Lan3-2-2	Western	3.8	<2.3	2,705 ± 7	170 ± 0.3	6.51 ± 0.02	1.004 ± 0.005	91.0 ± 1.9	249.6 ± 1.6	185 ± 4
Lan3-4-1	Western	2.3	<2.3	3,001 ± 8	110 ± 0.2	3.78 ± 0.01	0.996 ± 0.005	83.8 ± 1.9	249.6 ± 1.6	170 ± 9
Lan1-0-1	Kapihaa headland	8	0	3,153 ± 9	25 ± 0.3	0.82 ± 0.01	0.923 ± 0.005	85.0 ± 5.1	196.2 ± 1.3	148 ± 9
Lan2-2-3	Kapihaa headland	4.2	<2.3	2,754 ± 7	209 ± 0.7	7.86 ± 0.03	0.997 ± 0.006	84.7 ± 2.0	250.0 ± 1.7	172 ± 4
Archeology gully										
Lan2-7-1	Low	9.7	0	2,838 ± 7	145 ± 2.5	5.28 ± 0.09	0.803 ± 0.004	111.7 ± 2.3	134.5 ± 0.9	164 ± 3
Lan2-8-1	Upper	35	3.3	3,068 ± 8	607 ± 2.1	20.4 ± 0.09	0.945 ± 0.006	82.4 ± 1.8	211.4 ± 1.5	150 ± 3
Kapihaa gully										
Lan1-2h-1	Low	20	2.3	3,043 ± 8	637 ± 2.7	21.6 ± 0.11	0.783 ± 0.009	102.0 ± 3.2	130.9 ± 1.5	148 ± 5
Lan3-6-2	Low	19	0	2,741 ± 7	244 ± 0.5	9.20 ± 0.03	0.801 ± 0.004	108.2 ± 2.0	134.8 ± 0.9	159 ± 3
Lan2-3-1	Low	18	0	3,048 ± 8	98 ± 0.2	3.31 ± 0.01	0.985 ± 0.005	81.2 ± 1.9	242.6 ± 1.6	161 ± 4
Lan3-5-3	Mid	23.0	0	3,468 ± 9	1,256 ± 2.3	37.4 ± 0.12	0.798 ± 0.004	99.5 ± 1.4	136.4 ± 0.9	146 ± 2
Lan2-4-3	Mid	19	<2.3	3,071 ± 8	739 ± 1.1	24.9 ± 0.07	0.800 ± 0.005	101.4 ± 1.2	136.7 ± 1.0	149 ± 2
Lan1-4-1	Mid	28	2.3	3,133 ± 8	582 ± 3.7	19.2 ± 0.13	0.967 ± 0.016	78.4 ± 3.6	230.2 ± 3.9	151 ± 7
Lan3-5-1	Mid	17.7	2.3	3,094 ± 8	1,241 ± 2.1	41.4 ± 0.13	0.932 ± 0.005	81.3 ± 2.0	204.1 ± 1.4	145 ± 4
Lan1-6-1	Upper	58	0	2,896 ± 8	151 ± 0.5	5.39 ± 0.02	0.965 ± 0.005	86.3 ± 4.1	222.4 ± 1.5	162 ± 8
Lan2-6-2	Upper	58.9	0	3,008 ± 8	18 ± 0.3	0.62 ± 0.010	0.980 ± 0.005	83.7 ± 2.1	235.9 ± 1.5	163 ± 4
Lan2-5-2	Upper	47	<2.3	2,849 ± 7	12 ± 0.4	0.44 ± 0.02	1.005 ± 0.021	91.3 ± 2.3	250.0 ± 5.2	185 ± 5
Lan3-7-1	Upper	58	0	2,745 ± 8	1,029 ± 1.0	38.7 ± 0.05	1.006 ± 0.005	85.2 ± 1.9	258.2 ± 1.0	177 ± 3

Activity ratios were calculated using $\lambda^{238}\text{U} = 1.551 \times 10^{-10} \text{ yr}^{-1}$, $\lambda^{234}\text{U} = 2.835 \times 10^{-6} \text{ yr}^{-1}$ and $\lambda^{230}\text{Th} = 9.195 \times 10^{-6} \text{ yr}^{-1}$. U isotopic compositions are reported using the $\delta^{234}\text{U}$ notation, where $\delta^{234}\text{U} = [(^{234}\text{U}/^{238}\text{U})_{\text{measured}} - 1] \times 1,000$; errors in Th and U concentrations are based on replicates and standards analysis³². Errors on $\delta^{234}\text{U}$ ratios are based on $^{234}\text{U}/^{238}\text{U}$ analysis errors. ($^{230}\text{Th}/^{238}\text{U}$) errors are based on $^{230}\text{Th}/^{238}\text{U}$ analysis errors, and errors on Th and U concentration. Fractional errors in all of these parameters were propagated through the appropriate equations to estimate errors in ages and $\delta^{234}\text{U}$. Reported uncertainties in activity ratios and ages also include errors in λ values.

Analyses were conducted at the University of Hawaii by peak-jumping of masses 229–230–232 (Th) and 233–234–235 (U) in single-collector ion-counting mode using a Vacuum Generators Sector54 mass spectrometer outfitted with a Daly detector and a WARP (wide area reducing potential) secondary energy filter for improved abundance sensitivity. Data were corrected for Daly Bias of 0.58‰/a.m.u. relative; other instrument-related metadata are given elsewhere³². U isotopic fractionation was calibrated with external standards: Nine analyses of NIST certified reference material U010 yielded $^{234}\text{U}/^{238}\text{U} = 5.4637 \times 10^{-5} \pm 0.3\%$ (2σ relative), which is 0.03% lower than the certified value ($5.4655 \times 10^{-5} \pm 0.9\%$).

The remaining less reliable ages (243 to 258 kyr ago) are probably too old by ~15–30 kyr).

Ages of clasts are used here as indicators of general relationships and as tracers of source material to test depositional scenarios. Although the youngest clast limits the maximum age of a host stratum, clast ages are not deposit ages, as they may have had a finite age before their incorporation into a deposit, particularly if they have been reworked and redeposited from older strata. Ages in the LSG deposit in Hulopoe bay indicate that the time interval between coral death and subsequent incorporation into a potentially analogous deposit can be relatively short (2–3 kyr ago), but this may not always be the case.

Clast age distribution patterns

On the basis of our dating results, a single depositional event (such as the proposed giant tsunami) would require mobilization of a mixed-age assemblage of corals from one source (almost certainly the bay floor fronting the deposits). Age relationships within the MB (that is, material unconformably mantled by the coastal terrace) can be used to discriminate single-deposit versus multiple-deposit scenarios. Rationally, a single-depositional event (a giant tsunami) should not produce statistically significant clast age sorting and should be distinguished from non-catastrophic, sequential deposition of strata by lack or presence of such sorting.

We used standard probability and statistical hypothesis-testing methods to investigate whether the observed clast-age distribution could have been derived from a single clast assemblage. Clast distribution was examined using spatial domains defined by geography and stratigraphy within the MB. Calculated probabilities are not likely to have been biased by sampling deficiencies because the proportion of dated clasts relative to those collected from each of the spatial domains (33–56%) was not significantly different from the overall analysis rate of 42%. Overall, two-thirds of the Hulopoe gravel coral clasts are stage 7 (broadly) and one-third are stage 5e (regardless of whether the coastal terrace samples are included). We can use this as a source clast population or assume equal numbers of both ages and the result is the same, so long as the source clast pool is large.

Strong age sorting trends by height in the section and relative elevation of deposition are observed. There is a natural break in ages (Fig. 1a and b) distributed throughout the gravels at about 2 m height from the gully floors (<2 m, 90% of the clasts are stage 7, 10% are stage 5e; >2 m elevation, 100% are stage 5e). The probability of this sorting arrangement from a large pool of stages 7 and 5e clasts is 1 in 10,000 (0.01%). Within the present sample set, no

stage 5e samples were collected over 30 m in elevation. There is again less than 0.01% probability that a single event could have divided clasts into the observed populations of >30 m (100% stage 7) and <30 m (56% stage 5e, 44% stage 7). If the clast ages are separated into four populations (on the basis of height from gully floor, above/below 2 m, and elevation, above/below 30 m), the probability that the observed sorting was derived from a single event diminishes to 3 in 100,000.

The number of dated samples is relatively small, but we tested the statistical significance of these sample populations using the null overlap hypothesis for mean ages and variances with the Student's *t*-distribution for small sets (95% confidence limit). Our results demonstrate that these probabilities are derived from valid distinct populations. Mean ages in the sample groups <2 m and >2 m from gully floor, and in the sample groups <30 m elevation, >2 m from base and >30 m elevation, <2 m from base could not have been drawn from the same original sample population because their upper and lower *t*-distribution limits do not overlap. Thus both height in section and combined height in section plus depositional elevation are robust groupings for deposit discrimination using clast ages. By contrast, mean ages of samples in the >30 m and <30 m groups do overlap at the 95% confidence limit, indicating that depositional elevation alone is not a robust discriminant.

We conclude that the degree of stratigraphically controlled clast age sorting is incompatible with deposition of the entire package of gravels in a single event (tsunami or otherwise), invalidating the main premise of the giant wave hypothesis. On the basis of the 95% confidence level statistics, additional clast ages are unlikely to change the observed sorting, and thus the multi-event result (although they might extend the geographical coverage of the stage 5e strata). The geological significance of this clast age distribution is that the 2 m horizon within the MB correlates with the top of the basal units in the lower-elevation portions of the gullies studied. No datable samples were recovered from the intervening alluvial stratum, yet all samples at higher MB stratigraphic levels are stage 5e aged. Thus clasts in strata above and below 2 m were derived from fundamentally different source material (specifically, stage 5e versus stage 7 reef). The change in source materials aged ~70 kyr apart suggests a considerable time break between deposition of the basal layers (followed by erosion, soil formation, alluvial deposition and additional soil formation¹³) and the stage 5e layers. Only stage 7 samples were found in the upper elevation outcrops. Their stratigraphic relationship to the lower elevation basal units is unclear but they may be the source beds of stage 7 clasts (via down slope reworking after their original deposition).

Formation of the Hulopoe gravel

Besides that of the giant wave hypothesis, other studies have used geochronology to explain subaerial exposures of bioclastic sediments found throughout Hawaii^{14,15,20,21}. Late Quaternary uplift in the high central Hawaiian Islands was used to rationalize now-emergent *in situ* reef deposits on O'ahu^{14,15}. Also, uplift was proposed²¹ from a broad range of electron-spin resonance (ESR) ages of coral clasts at 5–30 m elevation on O'ahu (9 at 122 to 562 kyr ago) and 10–70 m elevation on Moloka'i and Lana'i (27 at 170 to 303 kyr ago, 2 at 140 to 146 kyr ago, and 1 at 37 kyr ago). Clast ages in both previous Lana'i studies^{4,21} were determined by relatively imprecise methods and were not meaningfully tied to the geologic origins of the beds where they were sampled. Although ages differ, neither previous study documented an age progression through the Hulopoe gravel²¹.

Other authors have also discussed whether the giant wave could have formed all or part of the Hulopoe gravel^{2,4,10,13,21–23}. Even before the present data were obtained there were difficulties with the giant wave model: analogous modern deposits have not been observed (most tsunami deposits are significantly thinner, containing primarily sand-sized grains and few boulders²⁴); the Hulopoe gravel

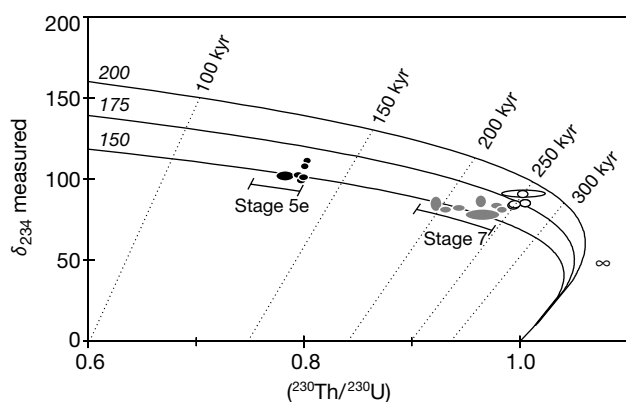


Figure 2 ^{230}Th – ^{234}U – ^{238}U data for Hulopoe gravel coral clasts. Closed-system evolution curves with different initial $^{234}\text{U}/^{238}\text{U}$ (given as δ values of 150, 175 and 200) and isochrons through the data field (dashed lines) are also shown. Data are plotted as ellipses centred on the data points and extending in the *x*- and *y*-directions to encompass 2σ analytical errors in ($^{230}\text{Th}/^{238}\text{U}$) and $\delta^{234}\text{U}$. Symbol fill patterns as in Fig. 1.

contains numerous different deposition facies; tsunami models do not predict such large run-up heights from likely Hawaiian landslides; and the volume of living and dead coral just offshore of Lana'i is a small fraction of that required to instantaneously form a similar deposit today. There are also two ways to explain clast size observations²³ in the Moloka'i gravels (up-gully diminishing-mean coral clast sizes and abundances at constant basalt clast sizes and at increasing abundances). One explanation is that clast size variations reflect coral mobilization from a point source (an offshore reef) and "distal" fining as a tsunami wave decelerated across the landscape, leaving basalt clasts unsorted because they were not point-source-derived²³. But this sorting can also be rationalized in a second way, by a variable-energy depositional environment (for example, approaching a beach from offshore), wherein the less robust coral clasts are preferentially fragmented, resulting in diminished sizes and abundance up the energy gradient, without affecting basalt clasts. In fact, it is difficult to envisage how the tsunami mechanism would size-sort corals and yet not sort denser basalt clasts of the same size modes.

We recognize the significance of large mass-wasting events in

shaping the Hawaiian landscape (see citations in refs 8, 10, 21) and the probability that such events sometimes result in tsunamis. However, the difficulties noted above and our data, which require several depositional events, oblige us to take a uniformitarian approach and reconsider the geologically plausible explanation that the Hulopoe gravel resulted primarily from more benign littoral deposition and relative sea-level change, as originally proposed^{1,3}. Such a scenario allows for large storms or normal tsunamis to have deposited local pockets of high-energy strata in the gravels.

Potential histories of the Hulopoe gravel should allow for two depositional events separated by ≤ 70 kyr (that is, the clast age differences between the basal unit and stratigraphically higher gravels in lower Kapihaa and Archeology gullies), as well as a second time gap of undetermined length preceding coastal terrace deposition. These histories should also consider the modern shoreline depositional environment, which has unconsolidated gravels with clasts of similar size, shape and type populations at the mouths of most gullies, extending to ~ 4 m above and to several metres below mean sea level. The proximal LSG deposit at about the modern mean sea level (composed of Holocene corals, dated here

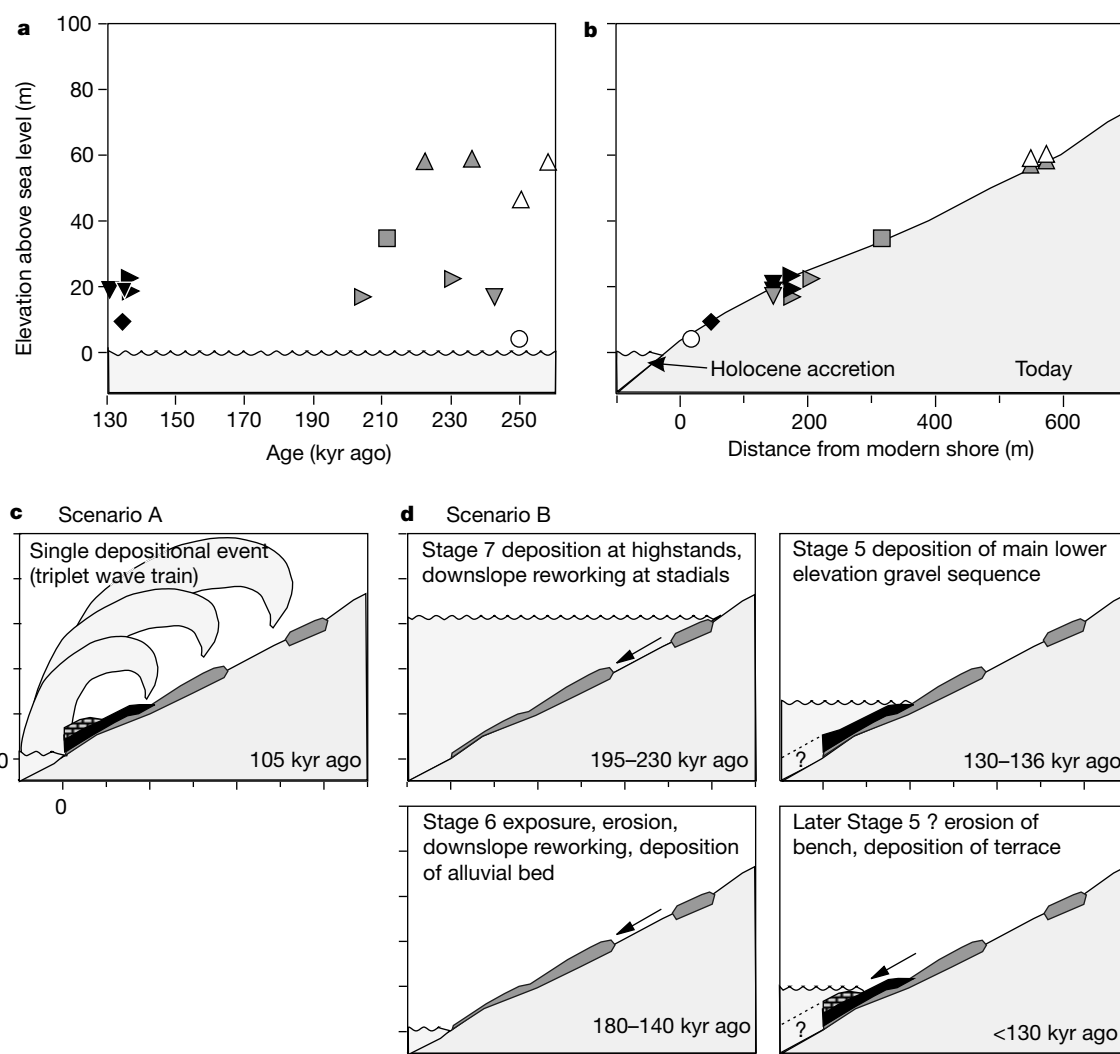


Figure 3 Age-elevation relationships for Hulopoe gravel coral clasts. **a**, Plot of age and present elevation above sea level. **b**, Composite section of Kapihaa and Archeology gully samples superimposed on the cross-sectional topography (vertical exaggeration is 5 times). Symbol types and fill patterns as in Fig. 1. The coastal terrace is shown as white bricks. Two scenarios (A and B) for forming the deposits are also shown in **c** and **d** (outcrop thicknesses have been exaggerated vertically by 100% to make them easier to

see; axes are as in **b**). Scenario A is the giant tsunami hypothesis. Scenario B involves sea-level variations and recent uplift of Lana'i. Arrows depict downslope reworking of the deposits. Only relative sea levels are given, as this scenario takes place in the context of both rising/falling sea levels and uplift of the island (the later of which has an unknown time–elevation pathway). Fill patterns for deposits in **b–d** are the same as in Fig. 1f.

at 2.5 ± 0.5 kyr ago, and older basalts) probably formed in an analogous environment to some Huloepoe gravel beds.

One scenario (Fig. 3) is that coral-bearing conglomerates were deposited, possibly over a range of upper and lower shoreface elevations, as sea level reached a maximum during the stage 7 deglaciation, and subsequently lithified. Like today, the stage 7 shoreline deposition was probably localized within relief at pre-existing gully mouths and laid down as a thinner veneer on interfluvies. Our maximum mapped extent of coral-bearing conglomerate was 63 m (in Kapihaa gully), although it is possible that it extended further upslope in the past. As sea level retreated, some stage 7 material was re-worked down slope, accompanied by new deposition, to form the lower elevation basal layers. On the basis of the maximum elevation difference today between *in situ* stage 7 reef (-15 m)¹⁶ and clastic deposits (15 m) containing stage 7 corals at Kahe point, O'ahu²⁰, we would expect to find remnants of *in situ* reef at lower elevations in Kapihaa gully today if it had existed in the past. Either it has been completely eroded or a large fringing reef never existed.

Following sea-level retreat, the basal deposit was partially eroded, an alluvial deposit was emplaced, sea levels subsequently rose to the stage 5e interglacial maximum, and the next sequence of gravels were deposited, again primarily filling preexisting topography. These attain a maximum elevation of 25 m in our sample set, which is some 18 m higher than the *in situ* stage 5e reef on O'ahu today^{14,15}. After the maximum stage 5e interglacial, the terrace sequence was deposited over both gravel-filled gullies and bare basalt headlands. Clast type¹³ and age populations indicate that it probably contains upslope alluvium. The time gap is difficult to determine, but the terrace is probably also a stage 5 deposit, possibly a prograding bed from the 5e regression, as no younger clasts were found. The large proportion of sand-size marine fossils, abundant corals and macrofaunal shells, strong cementation and beach rock clasts are consistent with deposition in a marine littoral environment.

Although the Huloepoe gravel is consistent with transgressive–regressive sea-level cycles, its vertical extent is significantly greater than sea-level change over the period and clearly implies that the island has been uplifted. Taking into account the range of estimated global stage 7 and 5e sea levels (-6 m to $+9$ m and $+2$ to $+8$ m, respectively)^{15,25–27}, maximum elevations of stage 7 and 5e outcrops in Kapihaa gully today suggest mean uplift of 0.23 – 0.29 and 0.15 – 0.22 mm yr⁻¹, respectively. The ranges depend on uncertainty in maximum elevations, palaeo sea levels, and depositional depths. These rates are means derived from one time–elevation point but the data do not require that uplift be a smooth function in time. These rates are lower than previous estimates (0.33 mm yr⁻¹)²¹, are ~ 3 – 5 times that of O'ahu^{14,15,21}, and are at the low end of the Sumba Island, Indonesia, range (0.2 – 0.5 mm yr⁻¹; ref. 11).

There are a number of potential causes for Late Quaternary uplift of tectonically active Lana'i (which experienced one of the largest historical Hawaiian earthquakes, the $M = 7.0$ earthquake of 1871 (ref. 28)). We do not discuss them in detail here, but two viable mechanisms are lithospheric flexure of the Pacific plate in response to volcanic loading on the Big Island^{29,30} and isostatic rebound following mass redistribution in large landslides³¹. Isostatic rebound can cause uplift of 1 – 100 m using a range of plausible landslide volumes³¹. Likewise, uplift can be plausibly modelled at Lana'i and Moloka'i by lithospheric flexure within reasonable bounds on elastic plate thickness and on the local rheology of the lithosphere/asthenosphere, the time-dependence of volcanic loading on the Big Island, and the potential effect of the Molokai fracture zone (P. Wessel, personal communication). Notably, flexure models predict a steep increase in subsidence approaching the Big Island, such that it and parts of southern Maui can be subsiding while Lana'i, Moloka'i and O'ahu are uplifting. Thus, tidal gauge evidence originally used to dismiss emergent sea level deposits on Lana'i^{2,4} is

not necessarily applicable.

Although great landslides are a common part of the evolution of Hawaiian volcanoes and probably result in tsunamis of some sort, we here present evidence that the Huloepoe gravel was not deposited in a single event. This invalidates the basic premise of the original giant wave hypothesis^{2,4}. Rather than being the type deposit of a mega-tsunami with a 375-m run-up, we propose that coral clast ages, gravel stratigraphy, and facies variations within these gravels record a multi-event/multi-environment sequence of normal littoral and alluvial deposition. Additional work is required to fully parametrize the uplift and subsidence history of the Pacific plate around Hawaii, although the magnitude of uplift proposed here is not unreasonable. Despite the present analysis, controversy about other subaerial fossiliferous sediments in Hawaii might continue, as the sedimentary records are difficult to interpret in the field. Future study could focus on the less fully understood aspects of the proposed alternative model for deposition during eustatic sea-level variations and vertical movement of Lana'i, or on other plausible depositional scenarios consistent with the geological and geochronological observations. □

Methods

Sample chips (~ 2 g) were taken from coral clasts that appeared to be largely pristine in all or part of their interiors (discernible skeletal structure, visual lack of secondary precipitates and/or extensive discoloration). In all, 50 corals were sampled at 8 sites; 36 of these were fresh enough to screen for mineralogical purity. Chips were physically cleaned under a binocular microscope, ultrasonically washed/leached with ultra pure water and 0.05 M HNO_3 in a clean room, dried under filtered air, and powdered in a synthetic corundum mortar. Powdered splits were analysed for aragonite and calcite abundance by X-ray diffractometry. Calcite was fully resolved from baseline at 2.3% and only samples with $\leq 2.3\%$ calcite were analysed for U-series isotopes. One 'acceptability threshold' exception was made (Lan2-8-1, with 3.3% calcite) because landscape grading subsequently destroyed the sampling locale.

Th and U were separated and purified from the bulk sample by anion exchange methods³². Samples were weighed, dissolved in 0.5 M HNO_3 , centrifuged to remove any insoluble (non-carbonate) material, spiked with calibrated ^{229}Th and ^{233}U tracers, evaporated to dryness, fumed in ultra pure aqua regia, evaporated to dryness, and equilibrated and dissolved into 500 μl of 7.5 M HNO_3 for Th and U separation. Th and U concentration and isotopic composition analysis was by high precision thermal ionization mass spectrometry of samples loaded on Aquadag carbon on outgassed single Re filaments. Total procedural blanks were 1 – 5 pg U, 5 – 10 pg Th. U-series results are reported in Table 1 using parentheses to denote activity ratios. Data uncertainties are discussed in Table 1.

Ages were calculated using³³: $1 - (^{230}\text{Th}/^{238}\text{U}) = e^{-\lambda_{230}t} - (\delta_{234}/1,000)$ ($\lambda_{230}/(\lambda_{230} - \lambda_{234})$)($1 - e^{-(\lambda_{234} - \lambda_{230})t}$). A detrital Th correction¹⁵ was not applied because corrections assuming Th/U of a Hawaiian tholeiite as detritus are smaller than analytical errors. Age reliabilities were investigated with geochemical tests for closed-system isotopic evolution since coral death, including evidence of diagenesis (recrystallization to and/or secondary deposition of calcite), U content, Th/U ratio, $^{234}\text{U}/^{238}\text{U}$, and initial $^{234}\text{U}/^{238}\text{U}$ (calculated assuming closed-system evolution with: $\delta^{234}\text{U} = \delta^{234}\text{U}_i e^{-\lambda_{234}t}$)³³. Calcite content ranges from 0 to 3.3% but calculated ages and other U-series parameters are not correlated with it. U content ranges from 2.61 to 3.47 p.p.m.—comparable to other Hawaiian corals (for example, 2.3 to 3.4 p.p.m. for >30 last interglacial corals¹⁵). Th/U ratios are variable (4.4×10^{-6} to 4.1×10^{-4}) and may indicate some detrital Th addition or U loss (for example, Th/U in Atlantic Ocean sea water³⁴ is 3.7×10^{-5}), but even our highest Th/U value is well within the range of Th/U in pristine corals from localities worldwide^{26,27,33,35}, and calculated age does not correlate with Th content or Th/U.

Age-corrected $^{234}\text{U}/^{238}\text{U}$ in a coral should compare to that of modern sea water if the coral has remained chemically closed since death and if $^{234}\text{U}/^{238}\text{U}$ in sea water has remained constant since formation. $\delta^{234}\text{U}$ in modern corals and $\delta^{234}\text{U}_i$ in Holocene corals are typically indistinguishable from the modern seawater value (see refs 32 and 33; K.H.R., unpublished work). Elevated $\delta^{234}\text{U}_i$ observed in some older corals probably indicates open-system behaviour^{11,16,26,36} (although it has been suggested that $^{234}\text{U}/^{238}\text{U}$ in sea water may have differed in the past³⁶). Most workers apply a working definition of ^{230}Th – ^{234}U – ^{238}U age quality based on $\delta^{234}\text{U}_i$ relative to modern: 145 – 150 is considered highly reliable, 150 – 160 or 165 is moderately reliable, and $\delta^{234}\text{U}_i > 165$ is less reliable^{11,15,35}. These ranges are somewhat arbitrary because acceptable $\delta^{234}\text{U}_i$ varies with how and when open-system behaviour occurred and the absolute age of the sample, and because not all open-system events that modify $\delta^{234}\text{U}_i$ also affect sample age^{27,36}. All of the Holocene Lana'i samples, 4 of 6 stage 5e samples, and 3 of 12 stage 7 samples have modern $\delta^{234}\text{U}_i$ values; this supports an essentially constant $^{234}\text{U}/^{238}\text{U}$ in sea water over the past 250 kyr³⁷ and suggests that samples with elevated $\delta^{234}\text{U}_i$ have been variably compromised. There is a clear inverse correlation of age and U content in the Lana'i stage 7 samples, suggesting U loss from the oldest samples (most probably leached by meteoric water without associated calcite recrystallization, consistent with the arid environment and lack of significant ground water in the deposits today). Contemporaneous *in situ* redeposition of ^{230}Th may also have occurred²⁷. There is also an inverse correlation of $\delta^{234}\text{U}$ and $\delta^{234}\text{U}_i$

with U in the stage 7 samples, as well as a lesser but still discernible correlation in the stage 5 samples (although ages do not covary with U content in these). It is likely that ^{234}U was recently added in a second alteration event in proportionally greater amounts to samples with lower U concentrations, although this was not by wholesale modern marine U addition from abiogenic calcitic or aragonitic cements as this would result in a negative correlation of $\delta^{234}\text{U}$ and age, opposite to that observed.

Received 30 March; accepted 25 October 2000.

1. Stearns, H. T. Geology and ground-water resources of the islands of Lana'i and Kahoolawe, Hawaii. *Hawaii Division of Hydrography/US Geol. Survey Bull.* **6**, 1–177 (1940).
2. Moore, J. G. & Moore, G. W. Deposit from a giant wave on the island of Lana'i, Hawaii. *Science* **226**, 1312–1315 (1984).
3. Stearns, H. T. Quaternary shorelines in the Hawaiian Islands. *Bishop Museum Press Bull.* **237**, 57 (1978).
4. Moore, G. W. & Moore, J. G. Large-scale bedforms in boulder gravel produced by giant waves in Hawaii. *Geol. Soc. Am. Spec. Pap.* **229**, 101–110 (1988).
5. Moore, J. G. & Fornari, D. J. Drowned reefs as indicators of the rate of subsidence of the island of Hawaii. *J. Geol.* **92**, 752–759 (1984).
6. Moore, J. G. in *Volcanism in Hawaii* (eds Decker, R. W., Wright, T. L. & Stauffer, P. H.) Ch. 2, 85–100 (USGS Professional Paper 1350, Washington DC, 1987).
7. Nisbet, E. & Piper, D. J. W. Giant submarine landslides. *Nature* **392**, 329–330 (1998).
8. McMurtry, G. M. *et al.* Stratigraphic constraints on the timing and emplacement of the Alike 2 giant Hawaiian submarine landslide. *J. Volcanol. Geotherm. Res.* **94**, 35–58 (1999).
9. Young, R. W. & Bryant, E. A. Catastrophic wave erosion on the southeast coast of Australia: impact of the Lana'i tsunamis ca. 105ka? *Geology* **20**, 199–202 (1992).
10. Moore, J. G., Bryan, W. B. & Ludwig, K. R. Chaotic deposition by a giant wave, Molokai, Hawaii. *Geol. Soc. Am. Bull.* **106**, 962–967 (1994).
11. Bard, E. *et al.* Pleistocene sea levels and tectonic uplift based on dating of corals from Sumba Island, Indonesia. *Geophys. Res. Lett.* **23**, 1473–1476 (1996).
12. Rubin, K. H., Sherman, C. E. & Fletcher, C. H. III Ages of emerged coral deposits in Kapihaa Gulch, Lana'i, Hawaiian Islands and speculation about their environment of deposition. *Trans. Am. Geophys. Union* **76**, F307 (1995).
13. Felton, E. A., Crook, A. W. & Keating, B. W. The Hulopoe Gravel, Lana'i, Hawaii: New sedimentological data and their bearing on the "Giant Wave" (mega-tsunami) emplacement hypothesis. *Pure Appl. Geophys.* **157**, 1257–1284 (2000).
14. Muhs, D. R. & Szabo, B. J. New uranium-series ages of the Waimanalo Limestone, O'ahu, Hawaii: Implications for sea level during the last interglacial period. *Mar. Geol.* **118**, 315–326 (1994).
15. Szabo, B. J., Ludwig, K. R., Muhs, D. R. & Simmons, K. R. Thorium-230 ages of corals and duration of the last interglacial sea-level high stand on O'ahu, Hawaii. *Science* **266**, 93–96 (1994).
16. Sherman, C. E., Fletcher, C. H. & Rubin, K. H. Marine and meteoric diagenesis of Pleistocene carbonates from a nearshore submarine terrace, O'ahu, Hawaii. *J. Sedim. Res.* **69**, 1083–1097 (1999).
17. Fletcher, C. H. III & Jones, A. T. Sea-level highstand recorded in Holocene shoreline deposits on O'ahu, Hawaii. *J. Sedim. Res.* **66**, 632–641 (1996).
18. Henderson, G. M. & Slowey, N. C. Evidence from U-Th dating against Northern Hemisphere forcing of the penultimate deglaciation. *Nature* **404**, 61–66 (2000).
19. Bassinot, F. C. *et al.* The astronomical theory of climate and the age of the Brunhes-Matuyama magnetic reversal. *Earth Planet. Sci. Lett.* **126**, 91–108 (1994).
20. Brückner, H. & Radtke, U. Fossile strände und korallenbänke auf O'ahu, Hawaii. *Essener Geogr. Arb.* **17**, 291–308 (1989).
21. Grigg, R. W. & Jones, A. T. Uplift caused by lithospheric flexure in the Hawaiian Archipelago as revealed by elevated coral deposits. *Mar. Geol.* **141**, 11–25 (1997).
22. Johnson, C. & Mader, C. L. Modeling the 105 ka Lana'i tsunami. *Sci. Tsunam. Haz.* **12**, 33–38 (1994).
23. Moore, A. L. Landward fining in onshore gravel as evidence for a late Pleistocene tsunami on Molokai, Hawaii. *Geology* **28**, 247–250 (2000).
24. Dawson, A. G. & Shi, S. Tsunami Deposits. *Pure Appl. Geophys.* **157**, 875–897 (2000).
25. Harmon, R. S. *et al.* U-series and amino acid racemization geochronology of Bermuda: implications for eustatic sea level fluctuation over the past 250,000 years. *Paleogeogr. Paleoclimatol. Paleoecol.* **44**, 41–70 (1983).
26. Gallup, C. D., Edwards, R. L. & Johnson, R. G. The timing of high sea levels over the past 200,000 years. *Science* **263**, 796–800 (1994).
27. Chen, J. H., Curran, H. A., White, B. & Wasserburg, G. J. Precise chronology of the last interglacial period: ^{234}U - ^{230}Th data from fossil coral reefs in the Bahamas. *Geol. Soc. Am. Bull.* **103**, 82–97 (1991).
28. Cox, D. C. *The Lanai earthquake of February 1871* 50 (Univ. Hawaii Environmental Center, Honolulu, Hawaii, 1985).
29. Watts, A. B. & Ten Brink, U. S. Crustal structure, flexure, and subsidence history of the Hawaiian Islands. *J. Geophys. Res.* **94**, 10473–10500 (1989).
30. Wessel, P. & Keating, B. H. Temporal variations of flexural deformation in Hawaii. *J. Geophys. Res.* **99**, 2747–2756 (1994).
31. Smith, J. R. & Wessel, P. Isostatic consequences of giant landslides on the Hawaiian ridge. *Pure Appl. Geophys.* **157**, 1097–1114 (2000).
32. Rubin, K. H. Analysis of ^{232}Th - ^{230}Th in volcanic rocks: a comparison of thermal ionization mass spectrometry and other methodologies. *Chem. Geol.* (in the press).
33. Edwards, R. L., Chen, J. H. & Wasserburg, G. J. ^{238}U - ^{234}U - ^{230}Th - ^{232}Th systematics and the precise measurement of time over the past 500,000 years. *Earth Planet. Sci. Lett.* **81**, 175–192 (1986/87).
34. Chen, J. H., Edwards, R. L. & Wasserburg, G. J. ^{238}U , ^{234}U , and ^{232}Th in seawater. *Earth Planet. Sci. Lett.* **80**, 241–251 (1986).
35. Stirling, C. H., Esat, T. M., Lambeck, K. & McCulloch, M. T. Timing and duration of the last interglacial: evidence for a restricted interval of widespread coral reef growth. *Earth Planet. Sci. Lett.* **160**, 745–762 (1998).
36. Hamelin, B., Bard, E., Zindler, A. & Fairbanks, R. G. ^{234}U - ^{238}U mass spectrometry of corals: how accurate is the U-Th age of the last interglacial period. *Earth Planet. Sci. Lett.* **106**, 169–180 (1991).
37. Henderson, G. M., Cohen, A. S. & O'Nions, R. K. ^{234}U - ^{238}U ratios and ^{230}Th ages for Hateruma Atoll corals: implications for coral diagenesis and seawater ^{234}U - ^{238}U ratios. *Earth Planet. Sci. Lett.* **115**, 65–73 (1993).

Acknowledgements

We thank E. Grossman and M. Coyne for field assistance, K. Spencer for maintaining the UH Isotope Lab, J. Moore, W. Bryan, A. Felton, R. Grigg, C. Glenn, T. Jones, P. Wessel and B. Keating for enlivening discussions, the Manele Bay Hotel for access and D. Muhs for a review. This work was supported by the University of Hawaii, Seagrant, the US Geological Survey and the US National Science Foundation.

Correspondence and requests for materials should be addressed to K.H.R. (e-mail: krubin@soest.hawaii.edu).

Crystal structure of Rac1 in complex with the guanine nucleotide exchange region of Tiam1

David K. Worthylake*, Kent L. Rossman† & John Sunde*†‡

* Department of Pharmacology, † Department of Biochemistry and Biophysics, ‡ Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA

The principal guanine nucleotide exchange factors for Rho family G proteins contain tandem Dbl-homology (DH) and pleckstrin-homology (PH) domains that catalyse nucleotide exchange and the activation of G proteins. Here we have determined the crystal structure of the DH and PH domains of the T-lymphoma invasion and metastasis factor 1 (Tiam1) protein in complex with its cognate Rho family G protein, Rac1. The two switch regions of Rac1 are stabilized in conformations that disrupt both magnesium binding and guanine nucleotide interaction. The resulting cleft in Rac1 is devoid of nucleotide and highly exposed to solvent. The PH domain of Tiam1 does not contact Rac1, and the position and orientation of the PH domain is markedly altered relative to the structure of the uncomplexed, GTPase-free DH/PH element from Sos1. The Tiam1/Rac1 structure highlights the interactions that catalyse nucleotide exchange on Rho family G proteins, and illustrates structural determinants dictating specificity between individual Rho family members and their associated Dbl-related guanine nucleotide exchange factors.

The major class of guanine nucleotide exchange factors for Rho family G proteins invariably contain tandem DH and PH homology domains that are sufficient to catalyse nucleotide exchange and concomitant G-protein activation^{1,2}. GTP-bound Rho family members coordinate actin cytoskeletal rearrangements and transcriptional activation^{3,4} to regulate diverse cellular processes including motility, neurite outgrowth and axonal guidance, morphogenesis, cytokinesis, membrane trafficking and cellular proliferation^{3,4}. In addition to performing key roles in many normal cellular processes, Rho family G proteins are also implicated in malignant growth transformation^{5,6}.

The activation of G proteins, including members of the Rho family, is dependent on replacing bound GDP with GTP, which is facilitated by guanine nucleotide exchange factors (GEFs). The largest family of GEFs that are specific for Rho family G proteins all share an ~300-amino-acid span of sequence homology, which was characterized initially in Dbl—an oncoprotein isolated from a diffuse B-cell lymphoma^{2,7}. This defining segment contains two domains, an ~200-residue Dbl-homology (DH) domain which is always in tandem with a carboxy-terminally located, ~100 residue PH domain. The invariant association of DH and PH domains in Rho family GEFs implies a functional interdependence. Isolated DH domains are often sufficient for nucleotide exchange, as indicated by mutational analysis and *in vitro* nucleotide exchange experiments. However, the functions of associated PH domains are less well-defined and may include subcellular localization and intramolecular regulation of exchange activity. Outside the DH and PH domains, Dbl family proteins vary greatly in sequence and domain architecture^{1,2}. Dbl family members exhibit various specificities toward Rho family G proteins, ranging from exquisitely specific to promiscuous. For instance, Fgd1 and Tiam1 are specific activators of Cdc42 (ref. 8) and Rac1 (refs 9, 10), respectively, whereas Vav supports nucleotide exchange on RhoA, RhoG, Rac1 and Cdc42 (ref. 11).

Although the cellular processes regulated by members of the Ras family of G proteins are widely known and extensively studied, the same is not true for Rho family members. To understand the roles of the DH and PH domains of Dbl family members in the activation of Rho family GTPases, we have determined the crystal structure of the

DH/PH fragment of Tiam1 in complex with Rac1. Tiam1 was identified through its ability to induce an invasive phenotype in a normally non-invasive T-lymphoma cell line, and has subsequently been shown to produce experimental metastases in nude mice¹². Tiam1 is expressed in many tissues and is present in virtually all tumour cell lines that have been analysed¹³. Aside from the characteristic DH/PH region, Tiam1 contains several domains and targeting sequences, including a second PH domain necessary for agonist-regulated translocation of Tiam1 from cytosolic to membrane fractions¹⁴.

The structure reveals that Tiam1 interacts primarily with the two switch regions of Rac1 resulting in their displacement and remodeling. These changes in Rac1 block magnesium binding and subtly alter the nucleotide-binding cleft to favour GDP release. Furthermore, non-conserved residues in Rac1 that contact Tiam1 highlight the region between switches that is likely to be important for

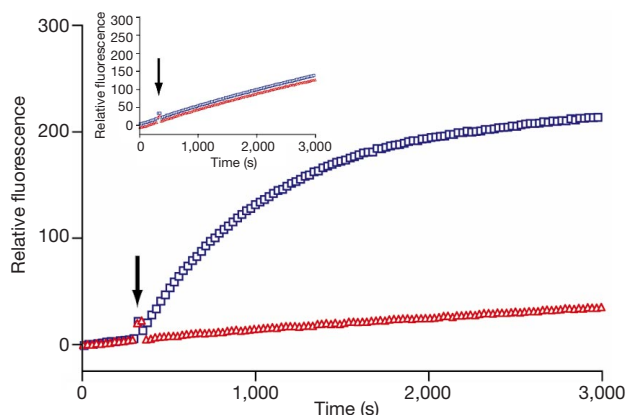


Figure 1 Tiam1 is a highly specific, functional GEF for Rac1 and not Cdc42. mant-GDP loading onto G proteins in the presence (blue, arrow indicates addition) or absence (red) of Tiam1 assayed by increased fluorescence. Tiam1 does not exchange guanine nucleotides on Cdc42 (inset). However, Cdc42 competently exchanges guanine nucleotides using several other GEFs (data not shown). Removal of residues 178–192 from full-length Rac1 has little effect on exchange (data not shown).

ensuring proper GEF/G-protein pairing. Finally, the PH domain does not contact Rac1, and its orientation with respect to the DH domain is distinctly different relative to the isolated DH/PH fragment of Sos1 (ref. 15), suggesting that either PH domains change conformation when GEFs bind GTPases, or there are many functions of DH-associated PH domains.

General features of the structure

Most GEFs interact strongly with nucleotide-depleted forms of their cognate G proteins, and we used this characteristic to isolate a fragment of Tiam1 in complex with nucleotide-free Rac1 (see Methods). Expressed and purified fragments of murine Tiam1 encompassing the DH and PH domains (1,033–1,406; 97.8% identical with human and hereafter referred to as Tiam1) and human

Rac1 (1–177; hereafter referred to as Rac1) were functionally characterized (Fig. 1) before crystallization in space group C_2 with four Tiam1/Rac1 complexes (each with a relative molecular mass (M_r) of 63,000 (63K)) in the asymmetric unit (see Methods). Initial phases were acquired using multiwavelength anomalous dispersion (MAD) from a single crystal grown using Tiam1 substituted with L-seleno-methionine and were improved by solvent flattening, histogram matching and 4-fold non-crystallographic symmetry (NCS) averaging using the program DM¹⁶ to produce a high-quality 3.0 Å experimental electron density map for use in model building (see Supplementary Information). The model was refined against all data with $|F| > 0$ extending to 2.8 Å; the final structure has a crystallographic R value of 26.2% (R_{free} 29.3%) and includes 2,065 residues, 93 water molecules and 4 sulphate ions.

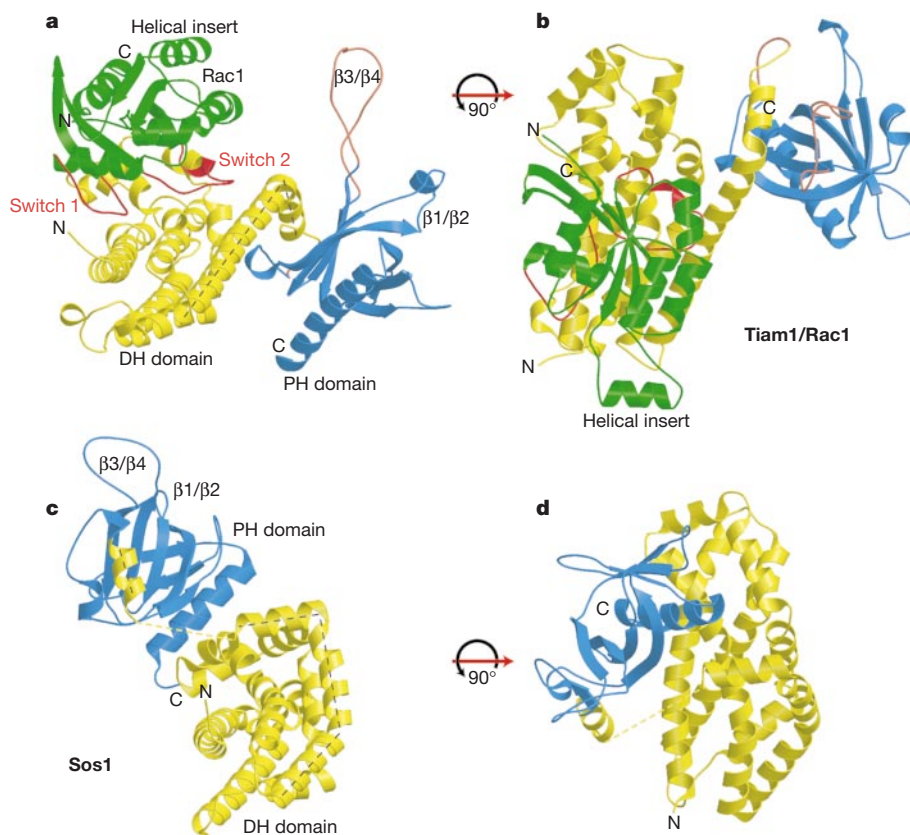


Figure 2 Structural comparison of the Tiam1/Rac1 complex with the Sos1 DH/PH domains. **a**, The Tiam1/Rac1 complex. Switch regions of Rac1 are red; the rest of the G protein is green. Tiam1 DH and PH domains are yellow and blue, respectively; disordered regions are orange; and black dashed line highlights $\alpha 9$ and its equivalent in Sos1. **b**, View in **a** rotated 90° about the horizontal. **c**, Structure of the Sos1 DH (yellow)

and PH (blue) domains¹⁵ aligned with Tiam1 orientated as in **a** by least squares superposition of the 72 α -carbon atoms of CR1–CR3 (r.m.s. deviations, 1.52 Å). Yellow dashed line indicates disordered residues in the Sos1 DH domain. **d**, View in **c** rotated 90° about the horizontal. Note, position of the Sos1 PH domain overlaps the G-protein-binding interface. Figures 2, 5 and 6c were made using MOLSCRIPT⁴⁰ and Raster3D⁴¹.

Table 1 Data, structure and refinement statistics

Data set	Wavelength (Å)	Resolution (shell) (Å)	Observations (Total/unique)	Completeness (%)	$\langle I/\sigma \rangle$ †	R_{sym} * (%)
λ_1	0.9792	15–2.8 (2.9–2.8)	382,085/80,441	94.2 (68.9)	27.9 (4.3)	6.9 (39.2)
λ_2	0.9787	15–3.2 (3.3–3.2)	388,784/66,980	99.3 (96.9)	29.4 (6.3)	7.6 (35.2)
λ_3	0.9724	15–3.0 (3.1–3.0)	405,551/68,882	98.9 (96.1)	30.9 (5.2)	7.0 (38.5)
Overall figure of merit‡						
Before density modification = 0.30						
After density modification = 0.75						
Data set	Resolution (Å)	Reflections (working/test)	Total atoms	$R_{\text{cryst}}/R_{\text{free}}§$	r.m.s.d. from ideal values	
					Bonds (Å)	Angles (°)
λ_1	15–2.8	76,460/4,051	16,164	0.262/0.293	0.007	1.42

* $R_{\text{sym}} = 100 \times \sum |I - \langle I \rangle| / \sum I$, where I is the integrated intensity of a measured reflection.

† $\langle I/\sigma \rangle$, Mean signal to noise, where I is the integrated intensity of a measured reflection and σ is the estimated error in the measurement.

‡ Figure of merit, $\langle |\Sigma P(\alpha)e^{i\alpha} / \Sigma P(\alpha)| \rangle$, where α is the structure factor phase and $P(\alpha)$ is the phase probability distribution.

§ $\Sigma |F_o - F_p(\text{calc})| / \Sigma F_p$, where F_o and $F_p(\text{calc})$ are the observed and calculated structure factor amplitudes. R_{free} is calculated similarly using test set reflections never used during refinement.

Numbers in parentheses pertain to the highest resolution shell. λ_1 data were collected first followed by λ_3 and then λ_2 .

The models for Rac1 are complete, but terminal residues of Tiam1 (1,033, 1,402–1,406) and a small region between the domains (1,253–1,258) are disordered and not included. In addition, the loops (1,306–1,330) connecting strands $\beta 3$ and $\beta 4$ in the PH domains are heterogeneously disordered and, together with several disordered side chains distant from the GEF/GTPase interface, were included during refinement with zero occupancy. The model has good geometry with overall r.m.s. differences from ideality in bond lengths and bond angles of 0.007 Å and 1.4°, respectively (Table 1). Because of a poor observation-to-parameter ratio we used tight NCS restraints throughout refinement, and three supplemental 5,000-degree simulated annealing refinements using torsion angle dynamics without NCS restraints indicate no significant structural differences between complexes (mean pairwise r.m.s. differences of all common C α atoms, 0.84 Å; maximum difference, 1.08 Å; minimum difference, 0.54 Å).

In the crystal asymmetric unit, Tiam1/Rac1 complexes dimerize to bury over 2,000 Å² of predominantly hydrophobic surface area mainly involving interactions between DH domains. The internal organization of dimers is consistent with simultaneous engagement of associated PH domains with lipid membranes^{17,18} and concomitant membrane insertion of the geranyl-geranyl modified C termini of Rac1, suggesting that Tiam1 may function as a dimer. Although both Ras-GRF1 (ref. 19) and Dbl (Y. Zheng, personal communication) are reported to dimerize through interactions mediated by DH domains, the functional implications of Tiam1 dimerization remain unclear. Tiam1 does not dimerize at micromolar concentrations in solution in the absence of phospholipids (our own unpublished data).

Tiam1/Rac1 complex

The overall structure of a single Tiam1/Rac1 complex is similar to a capital letter 'T', with the DH domain in the middle, and Rac1 and the PH domain forming the arms of the T (Fig. 2). The GEF/G-protein interface is extensive with over 3,000 Å² of predominantly hydrophobic, solvent-accessible surface buried on complex formation. Interactions between Rac1 and Tiam1 are mediated primarily by switches 1 (residues 25–39) and 2 (residues 57–75), and by Tiam1 residues located in and near conserved region (CR)-1 and

CR3 (see Fig. 3; and Supplementary Information).

The DH domain (residues 1,034–1,258) is an elongated helical bundle composed of nine α -helices and four 3_{10} -helices, and has the appearance of a 'chaise longue', with a long 'seat' comprising residues 1,034–1,058, 1,088–1,149 and 1,188–1,239, and a 'seat-back' formed by residues 1,066–1,087 and 1,150–1,187. The overall structure of the DH domain is very similar to the previously reported structures of the DH domains from Sos1 (ref. 15), β PIX²⁰ and Trio²¹ with two exceptions: first, helix $\alpha 9$ kinks at residue 1,240 to direct the C-terminal portion of $\alpha 9$ away from the body of the DH domain; and second, the Tiam1 seatback is in a more upright position relative to DH domains of Sos1 and β PIX.

The PH domain of Tiam1 is similar to other PH domains and forms an antiparallel β -sandwich, with sheets comprising strands $\beta 1$ –4 and $\beta 5$ –7, that is capped at one end by the C-terminal α -helix^{22,23} (Fig. 2). It is now apparent that DH-associated PH domains diverge from canonical structure within the region

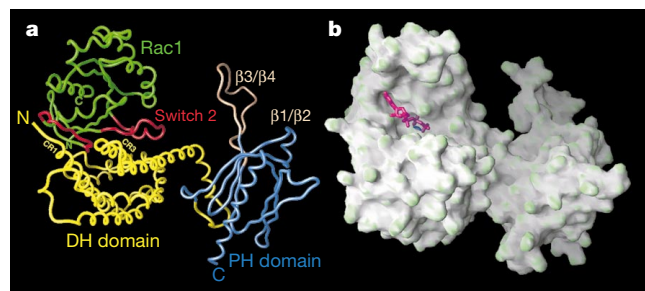


Figure 4 Surface representation of the Tiam1/Rac1 complex. **a**, Coil representation of the complex in roughly the same orientation as in Fig. 2a. **b**, A GRASP⁴³ surface representation of the Tiam1/Rac1 complex in the same orientation as in **a**, illustrating the accessibility of the nucleotide-binding cleft of Rac1 while bound to Tiam1. The GTP analogue (magenta) and Mg²⁺ (blue) were positioned by aligning the structure of Rac1/GMPNP onto Rac1 of the complex using a least squares superposition of α -carbons from residues 6–24, 52–56, 76–118 and 137–175 (r.m.s. deviations, 0.5 Å). The visibility of the base, ribose, all three phosphates and part of the Mg²⁺ illustrate the extensive solvent exposure of the active site.

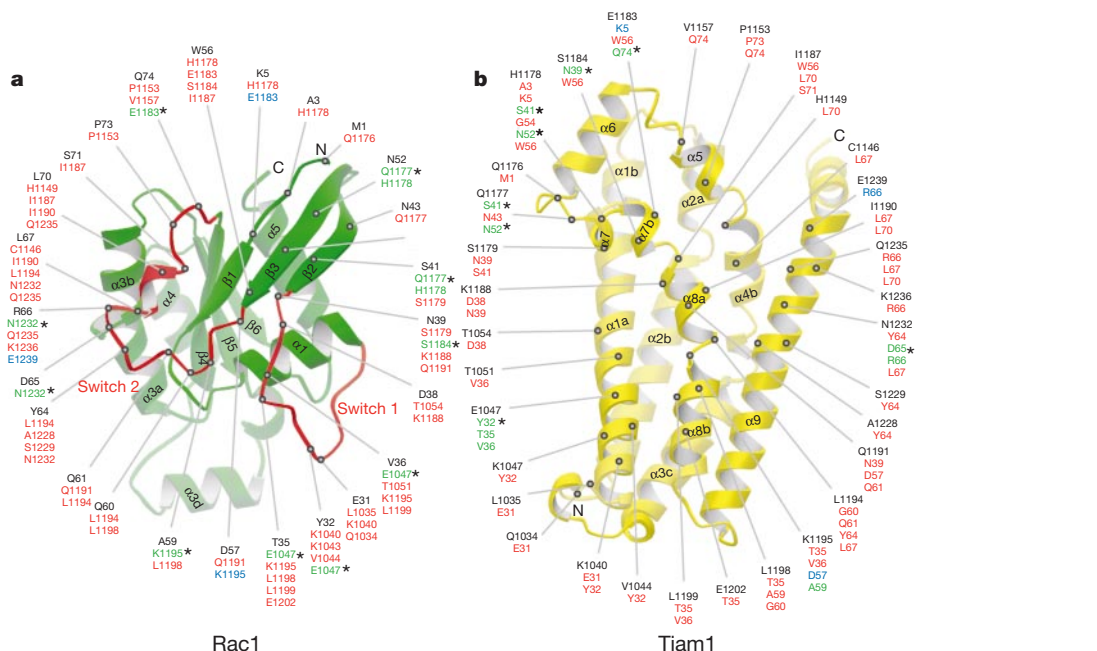


Figure 3 Representation of Tiam1/Rac1 interface contacts. The model of Rac1 on the left in green (with switch regions in red) has been rotated 180° about the vertical axis from its relative orientation with respect to the DH domain (on right in yellow) to display the two interface surfaces. Residues in black type lose >10 Å² of accessible surface area on

complex formation. Associated intermolecular contacts are listed below, colour coded by interaction type (red, van der Waals; green, hydrogen bonds; blue, ionic) and starred if non-ionic but involving side-chain atoms. Figure was made using Ribbons⁴².

immediately preceding the first β -strand. In Sos1 (ref. 15), this region contains an α -helix (α N), a β -strand (β N) and a 3_{10} -helix, and these elements pack mainly through van der Waals' interactions into the face of the β 1– β 4 sheet. In Tiam1, α N is incorporated into the C-terminal portion of the DH domain (α 9 residues 1,239–1,249), and there is no element analogous to β N of Sos1. However, the PH domain of Tiam1 does possess a 3_{10} -helix (residues 1,264–1,266) similar to that of Sos1, which participates with the β 1– β 4 sheet and the C terminus of the DH domain to bury 1,430 Å² of predominantly hydrophobic surface area.

The interaction with Tiam1 has altered the conformation of sequences in and immediately flanking the two switch regions of Rac1, but the remainder of the molecule is in essentially the same conformation as in Rac1/GMPPNP²⁴ and Cdc42/GDP (protein data bank accession number 1AN0, chain A; note that Cdc42 is 72% identical in sequence to Rac1 and there is currently no structure of GDP-bound Rac1). The P-loop of Rac1 (residues 10–17), which is integral for phosphate binding and is altered in other GEF/G protein structures²⁵, is left undisturbed and coordinates with a sulphate ion to reproduce interactions normally involving the β -phosphate of a bound guanine nucleotide. Rac1 sits on the seat of the DH domain, contacting CR1, CR3 and the C terminus of α 9, while presenting strands β 1–3 and part of switch 2 towards the seatback. The helical insert, which is unique to Rho family GTPases and implicated in effector regulation, is not contacting Tiam1. A surface representation of the complex indicates that the nucleotide-binding site is exposed to solvent and readily accessible to incoming guanine nucleotides (see Fig. 4).

Structural changes in switch regions

In the Tiam1/Rac1 complex, switch 1 is sandwiched between CR1 and CR3, and although both conserved regions make numerous contacts with Rac1, Glu 1,047 of CR1 mediates the most substantial series of interactions with switch 1. This residue participates in three hydrogen bonds with Rac1 involving the amide nitrogens of Thr 35 and Val 36, as well as the hydroxyl of Tyr 32 (see Fig. 5a). These interactions support the movement of Thr 35 of Rac1 $\sim$ 6 Å relative to its position in Rac1/GMPPNP ($\sim$ 2.4 Å from its position in Cdc42/GDP), thus preventing its ability to bind Mg²⁺. When compared with the Cdc42/GDP structure, the interaction of Tiam1 with Rac1 has shifted switch 1 up to 2.7 Å along the nucleotide-binding cleft. This shift in switch 1 places Ile 33 $\sim$ 2 Å from the position occupied by the ribose in the structures of Cdc42/GDP and Rac1/GMPPNP and would therefore sterically hinder nucleotide access to the binding pocket. In both the Rac1/GMPPNP and Cdc42/GDP structures, a hydrogen bond between the side chain of Cys 18 and the O2 atom of the α -phosphate is evident. Because of the Tiam1-induced shift of Ile 33, Cys 18 adopts a distinct rotamer and can no longer hydrogen bond with the nucleotide. Finally, Phe 28, which normally interacts favourably with the guanine base, is displaced $\sim$ 1 Å in the Tiam1/Rac1 complex, diminishing potential interactions with a bound nucleotide.

On complex formation, switch 2 buries $\sim$ 50% more solvent-accessible surface area than switch 1 (1,759 Å² compared with 1,193 Å²). In addition, switch 2 alterations are confined to residues 59–64 relative to Cdc42 bound to GDP (Fig. 5b). In Cdc42, this region begins as an extended conformation before terminating in a

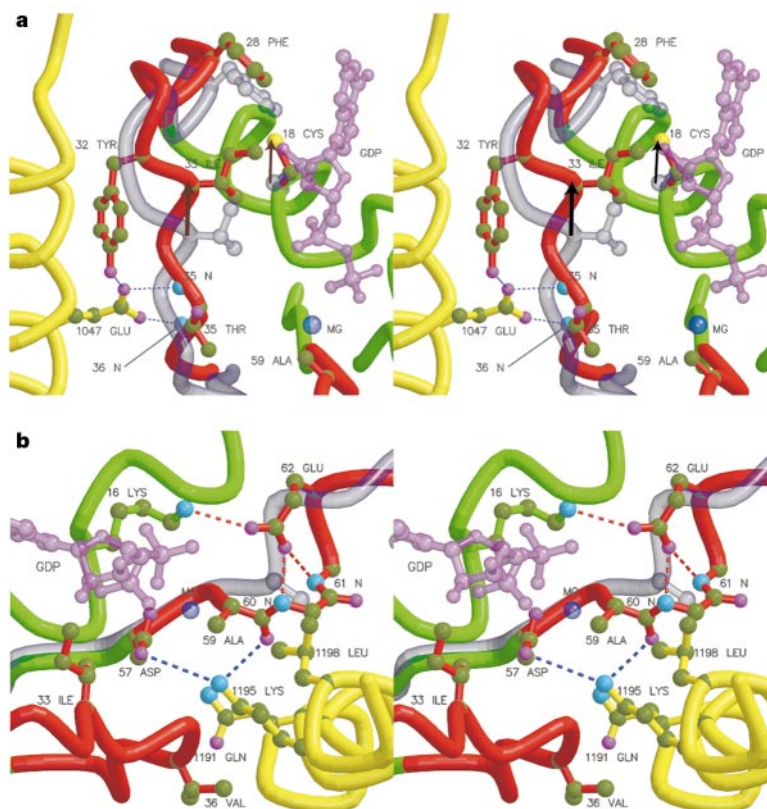


Figure 5 Stereo view of the switch regions of Rac1. **a**, Switch 1. Tiam1 DH domain residues are shown in yellow, and Rac1 is coloured green with the switch regions in red. Shown in transparent indigo is a coil representation of Cdc42/GDP residues 27–39, as well as the side-chain atoms of Val 33, Phe 28 and Cys 18. In transparent magenta and blue are the GDP and Mg²⁺ from the Cdc42/GDP structure. Note the hydrogen bonds made by Glu 1,047 of Tiam1, the displacement of switch 1 along the nucleotide-binding cleft (large arrow), and the alternate rotamer conformations of Cys 18 seen between the two structures (small arrow). At the lower right is Ala 59 of switch 2. The two structures

were aligned using residues described in Fig. 4 (r.m.s. deviations, 0.51 Å). **b**, The switch 2 region of Rac1. A coil representing Tiam1 residues 1,187–1,201 of conserved region 3, as well as the indicated Tiam1 side chains are shown in yellow. Tiam1, Rac1 and its switch regions are coloured as described in previous Figs. Intermolecular polar or ionic interactions are represented as blue dashed lines; intramolecular interactions are red dashed lines. Transparent GDP (magenta) and Mg²⁺ (blue) of Cdc42/GDP aligned as in panel **a**. This view highlights key GEF/G-protein interactions made by Lys 1,195 buttressed by Leu 1,198, Gln 1,191 of Tiam1 and Val 36 of Rac1.

short 3_{10} -helix (residues 61–65), but in Rac1 complexed to Tiam1, this region consists of two overlapping β -turns: 59–62 (type IV), 61–64 (type I). The α -carbon of Glu 62 in Rac1 moves ~ 4.7 Å relative to its location in Cdc42/GDP and the glutamate is no longer directed into solvent, but instead points toward the nucleotide-binding pocket and hydrogen bonds with amide nitrogens of both Gly 60 and Gln 61, while simultaneously forming an ion pair with Lys 16. The most important functional consequence of switch 2 alterations is the repositioning of Ala 59, which has moved to within ~ 2 Å of the Mg^{2+} -binding site to prohibit productive magnesium ligation critical for high-affinity nucleotide binding (Fig. 5). The rearrangement of Ala 59 is supported by Lys 1,195, which hydrogen bonds to the backbone oxygen of Ala 59; in other DH domains, the equivalent of Lys 1,195 is nearly invariant.

Regions proximal to residues 59–64 of switch 2 in Rac1 are shifted slightly but otherwise remain unperturbed on interaction with Tiam1. The burial of Tyr 64 of Rac1 into a hydrophobic pocket formed by both Tiam1 (Leu 1,194, Ala 1,228, Ser 1,229 and Asn 1,232) and Rac1 (Gln 61, Asp 63 and Leu 67) anchors switch 2, whereas amino acids 65–74 of Rac1 interact extensively with CR3 and flanking regions of Tiam1 to bury a total of 965 Å² of the solvent-accessible surface area. Amino-terminal to Ala 59 of Rac1, interactions are less extensive and primarily involve Asn 52 and Trp 56 (see Fig. 3).

Interpretation of DH domain mutations

The structure of Rac1 bound to Tiam1 provides a framework for interpreting existing mutational data that was designed to probe the mechanism of Dbl-related GEFs. For example, substitution of Ser 1,216 to phenylalanine in the DH domain of UNC-73 was originally identified as causing moderate uncoordinated movement in *Caenorhabditis elegans*; this mutation abrogates nucleotide exchange on Rac *in vitro*²⁶, and, with the exception of threonine, substitution of other amino acids at this site also severely impairs nucleotide exchange²⁰. The equivalent residue in Tiam1 (Thr 1,051) interacts with numerous DH domain residues and is apparently critical for maintaining structural integrity (Thr 1,051 to alanine in Tiam1 promotes protein unfolding; our own unpublished data). Of eight single or double mutations in the DH domain of Dbl²⁰, only the combined substitutions of Asn 673 and Asp 674 to alanine severely impaired nucleotide-exchange activity towards Cdc42. The equivalent of Asn 673 in Tiam1 (Asn 1,232) interacts with switch 2 residues (see Fig. 3), and its mutation should impair complex formation. For the remaining set of seven neutral mutations, equivalent positions in Tiam1 are solvent exposed, not in contact with Rac1, and should not affect nucleotide exchange on substitution.

On the basis of the three-dimensional structure of the Trio DH domain and NMR information localizing its Rac1-binding surface,

several site-directed mutations in the Trio DH domain were made²¹. Of the substitutions to alanine that significantly reduced ($> 20\%$ relative to wild type) the ability of the Trio DH/PH fragment to catalyse nucleotide exchange on Rac1, equivalent residues in Tiam1 are either implicated in maintaining the structural integrity of the DH domain (Thr 1,051, see above; Arg 1,201 and Arg 1,192 hydrogen bond with Asn 1,225 and Thr 1,051, respectively), or are intimately involved in creating the interface between Tiam1 and Rac1 (Glu 1,047, Ile 1,190, Gln 1,191, Leu 1,194, Lys 1,195 and Leu 1,198). In particular, the substitution of Lys 1,372 to alanine in the Trio DH domain is extremely deleterious towards exchange activity. Lys 1,372 of Trio most probably serves the same function as its equivalent in Tiam1 (Lys 1,195) and its substitution may affect the ability of Trio to orientate Ala 59 of Rac1 into the Mg^{2+} -binding pocket. (Note, several similar mutations have been made in Tiam1 (Leu 1,194 to Ala, Lys 1,195 to Ala and Leu 1,198 to Ala) and shown to reduce nucleotide-exchange activity to less than 20% of wild type (our own unpublished data).)

GEF/GTPase-binding specificity and exchange mechanism

Out of all the Rac1 residues that lose surface exposure on complex formation, nine positions (Ala 3, Glu 31, Asp 38, Ser 41, Asn 43, Asn 52, Gly 54, Trp 56 and Gln 74) are variable between RhoA, Rac1 and Cdc42 and participate with Tiam1 to bury 937 Å² of solvent-accessible surface area. Of this set, only positions 3, 41, 43, 52 and 56 show sequence variability between Rac1 and Cdc42, and most of these residues cluster in G-protein strands $\beta 2$ –3 to strongly implicate this region as a principal determinant dictating selective coupling of G proteins with DH domains (see Supplementary Information). Additionally, Tiam1 residues (1,177–1,179) that interact with strands $\beta 2$ –3 of Rac1 are among the most variable positions in other DH domains (see Fig. 6), suggesting a degree of correlated mutations within G proteins and GEFs necessitated by functional interaction.

It seems plausible that Tiam1 first interacts with conformationally rigid portions of Rac1 (strands $\beta 2$ –3 and residues 65–74 of switch 2) to provide sufficient binding energy for subsequent catalysis. This 'lock and key' interaction is followed by an 'induced fit' that results in alteration of the conformations of switch 1 and the remainder of switch 2 to destabilize nucleotide binding. During the induced fit, the combined effect of the intrusion of Ala 59 into the Mg^{2+} -binding site and the steric clash of Ile 33 of switch 1 with the ribose promotes GDP dissociation. The complex allows unimpeded access for subsequent binding of GTP and concomitant restructuring of the switch regions to conformations seen in the Rac1/GMPPNP structure. Restructuring of the switch regions ultimately destabilizes the Tiam1/Rac1/GTP ternary complex leading to dissociation of active Rac1/GTP from Tiam1.

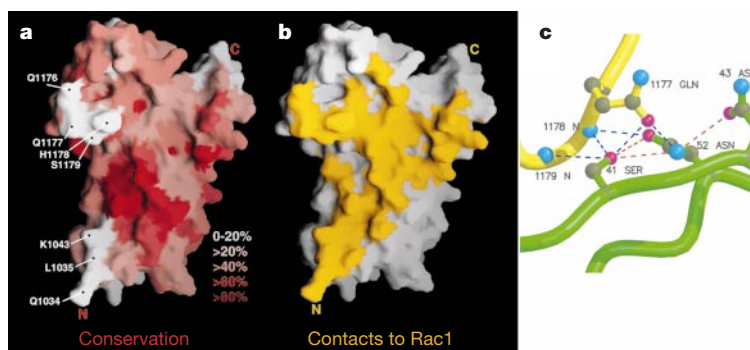


Figure 6 Interactions mediating proper GEF/G-protein pairing. **a**, Sequence conservation of 60 DH domains as determined by ClustalX has been mapped onto a GRASP⁴³ surface representation of the Tiam1 DH domain. Residues that contact Rac1 and are located in regions of low sequence conservation are shown. **b**, The molecular surface associated with DH domain residues that bury at least 10 Å² of solvent accessible surface area in the

Tiam1/Rac1 complex are yellow. **c**, Three of the five residues of Rac1 that contact Tiam1 and differ in sequence from Cdc42 are shown in green. In Cdc42, which fails to bind to Tiam1 *in vitro*, these residues are Ala 41, Thr 43 and Thr 52. All possible polar interactions less than 3.4 Å are indicated as blue (intermolecular) or red (intramolecular) dashed lines.

Comparison with other GEF/G-protein structures

Tiam1/Rac1 is the fourth GEF/G-protein structure to be elucidated, and was preceded by the structures of EF-Ts/EF-Tu²⁷, Sos1(Cdc25 domain)/Ras²⁸ and by Gea2(Sec7 domain)/Arf1 (ref. 29). In general, each of these complexes disrupts Mg²⁺ binding by the G protein and interferes with the binding of either the α - or β -phosphate to effect dissociation of bound nucleotide. Although the Tiam1/Rac1 structure features interference with Mg²⁺ binding, the structure differs from other GEF/G-protein complexes in that, with the exception of the induced alternative rotamer conformation of Cys 18 and the Lys 16/Glu 62 ion pair, it does not reveal direct interference with the binding of either the α - or β -phosphate of a guanine nucleotide. It also appears that the Tiam1/Rac1 structure represents the first instance of interference with the nucleotide sugar to assist in ejection of bound nucleotide.

In the Tiam1/Rac1 complex, residues 57–63 of Rac1 possess a conformation remarkably similar to the equivalent region of Ras in the structure of Sos1(Cdc25 domain)/Ras²⁸ and the Mg²⁺-free structure of RhoA/GDP³⁰; the three regions superimpose (pairwise) with an r.m.s. difference of ~ 1.0 Å for all common atoms. In fact, all three structures feature an ionic interaction between G-protein residues Lys 16 and Glu 62, hydrogen bonding between Glu 62 and the amide nitrogen of Gly 60, and occlusion of the magnesium-binding site by Ala 59 (in RhoA, Lys 18, Glu 64, Gly 62 and Ala 61, respectively). The two GEF/G-protein complexes share other common features including a glutamate/arginine salt-bridge C-terminal to Ala 59 (Glu 1,239/Arg 66 in Tiam1/Rac1, Glu 1,002/Arg 68 in Sos1(Cdc25)/Ras), presentation of a leucine side chain toward the Mg²⁺-binding site that packs against Ala 59 (Leu 1,198 and Leu 938 in Tiam1 and Sos1(Cdc25), respectively), and perhaps most importantly, intermolecular hydrogen bonding to the carbonyl oxygen of Ala 59 whose methyl group interferes with Mg²⁺ binding (by Lys 1,195 in Tiam1, Thr 935 in Sos1(Cdc25)). The Mg²⁺-free structure of RhoA/GDP indicates that residues 57–63 of switch 2 have an intrinsic propensity for adopting the conformation seen in Sos1(Cdc25 domain)/Ras and Tiam1/Rac1, whereas GEF/G-protein complex formation stabilizes this conformation in the presence of intracellular Mg²⁺ (~ 1 mM).

Guanine nucleotide dissociation inhibitors (GDIs) diametrically oppose GEF function by stabilizing GDP-bound G proteins³¹, and the structure of RhoGDI bound to Cdc42/GDP³², in comparison with Tiam1/Rac1, indicates significant overlap between GDIs and GEFs for Rho-family G-protein binding, particularly in the switch regions. This overlap strongly suggests preclusion of a stable RhoGEF/GTPase/RhoGDI ternary complex and also intimates at the need for the regulated removal of GDIs from G proteins, possibly by GEFs, before productive GEF interaction. Intriguingly, RhoGDI hydrogen bonds to the backbone nitrogen of Ala 59, and, in direct contrast to Tiam1, prevents intrusion of Ala 59 into the Mg²⁺-binding site.

Role of the PH domain

Evidence suggests that DH-associated PH domains have several functions, including localization of Dbl-family proteins to plasma membranes and intramolecular regulation of exchange activity through the binding of specific membrane components³³. Although the structure of Tiam1/Rac1 does not preclude interdomain regulation of exchange activity dependent on phosphoinositide binding to the PH domain, no relevant mechanism is obvious, and we have been unable to observe significant effects of phosphoinositides on Tiam1-catalysed nucleotide exchange *in vitro* (unpublished data). However, PH domains also directly regulate exchange activity independent of phosphoinositide binding. For example, relative to the isolated DH domain, a DH/PH fragment of Trio is ~ 100 -fold better at catalysing nucleotide exchange on Rac1 (ref. 21). In the Tiam1/Rac1 complex, C-terminal portions of the DH domain interact with both Rac1 and the PH domain, indicating that

structural stabilization of Tiam1 may be dependent on both the DH and PH domains and is necessary for full exchange activity. Further supporting a structural interdependence of DH and PH domains, the solution structures of the isolated DH domains of β PIX and Trio each terminate in highly disordered regions^{20,21}. In contrast, the analogous region of Tiam1 ($\alpha 9$) is well ordered and interacts with the PH domain (see Fig. 2).

In comparison to Sos1, the DH and PH domains of Tiam1 possess markedly altered relative orientations (Fig. 2), which allow unencumbered access of Rac1 to its binding surface on Tiam1. Although the structure of Tiam1 in the absence of Rac1 is not available, the efficient *in vitro* exchange data (Fig. 1) suggest the complex structure may provide a faithful representation of Tiam1 before Rac1 interaction. Furthermore, for Tiam1 to adopt a conformation resembling the DH and PH domains of Sos1 it would require disrupting a large interface, unravelling of $\alpha 9$, a traverse of ~ 55 Å by the PH domain, and the establishment of another large and similarly complementary interface. It should be noted that not only does Sos1 not support nucleotide exchange *in vitro* (S. Soisson, personal communication), but, relative to equivalent regions in other DH/PH sequences, the 'linker' region between the DH and PH domains of Sos1 is poorly conserved, possibly allowing unique conformations indicative of unusual function or regulation.

Conclusions

For all G proteins, the fundamental need to convey binary information as a function of bound guanine nucleotide produces common regulatory constraints. For example, all G proteins are tightly regulated against spurious activation and rely on GEFs to catalyse the exchange of GTP for bound GDP to initiate signalling cascades. But although all G proteins appear to share common ancestry, distinct G-protein families rely on seemingly unrelated families of GEFs. The structure of Tiam1 bound to Rac1 highlights similarities with other GEF/G-protein complexes, such as the requirement to destabilize Mg²⁺ interactions, but also details functional characteristics unique to the Dbl family of GEFs, such as the steric hindrance of the ribose sugar and the lack of any direct imposition of Tiam1 residues into the active site. The Tiam1/Rac1 structure also emphasizes the importance of the invariantly associated PH domain in maintaining the structural integrity of the DH domain and places constraints on how lipid binding to PH domains can affect nucleotide exchange. Finally, the large family of Rho proteins couple to an even larger family of Dbl-related GEFs, and the structure of Tiam1/Rac1 provides useful insights into how Dbl-family GEFs discriminate among G proteins. □

Methods

Biochemical characterization of Tiam1 and Rac1

The methionine auxotrophic *Escherichia coli* strain B834 (DE3) carrying a pProEX HT expression vector (Life Technologies, Grand Island, NY) was used to express selenomethionylated murine Tiam1 (residues 1,033–1,406) as a fusion protein with an N-terminal histidine tag. Five millilitres of cells grown overnight in LB were used to inoculate each litre of modified minimal media consisting of M9 supplemented with 42 mM glucose, 3 mM of each amino acid except methionine, 10 ml of 100 $\times$ MEM vitamins (Sigma) and 100 mg l-seleno-methionine (Sigma). Cells were grown to an optical absorbance at 600 nm (A_{600}) of 1.0 and expression of Tiam1 was induced with 1 mM IPTG for > 5 h at 37°C. Cells were collected by centrifugation and stored at -80°C . Thawed cells were resuspended in 20 mM MES pH 6.0, 50 mM NaCl, 5 mM EDTA, 5 mM dithiothreitol (DTT) (buffer A) and lysed by passage through a French press (Aminco) at 15,000 p.s.i. Cell debris was removed by centrifugation and the supernatant was loaded onto a 150-ml Fast Flow S-Sepharose column (Pharmacia, Piscataway, NJ) equilibrated to 20 mM MES pH 6.0, 50 mM NaCl, and proteins were eluted with a linear gradient in NaCl. Fractions containing Tiam1 were identified by SDS-polyacrylamide gel electrophoresis (PAGE), pooled, adjusted to a pH > 7.0 using 2 M HEPES pH 7.5 and loaded onto a 5-ml nickel-charged, metal-chelating column (Pharmacia) equilibrated to 20 mM sodium phosphate pH 7.5, 5% (v/v) glycerol and 100 mM NaCl (buffer B) containing 10 mM imidazole. Tiam1 was eluted with buffer B containing 300 mM imidazole and dialysed overnight against 20 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM EDTA, 5 mM DTT (buffer C) at 4°C with added tobacco etch virus protease to remove the histidine tag. Cleaved Tiam1 was concentrated to < 10 ml using a stirred cell concentrator (Amicon) and loaded onto a preparative grade, S-200 gel-exclusion column (Pharmacia; 26 mM/60 mM) equilibrated

to buffer C. Fractions containing Tiam1 were analysed by SDS–PAGE, pooled, dialysed into buffer A and loaded onto an 8-ml Source S column (Pharmacia), equilibrated to buffer A before elution with a gradient of NaCl. We identified fractions containing pure Tiam1 by SDS–PAGE.

A pET-15b plasmid vector encoding a C-terminal truncation mutant of human Rac1 (residues 1–177; a gift from S. Campbell, UNC Chapel Hill) was expressed in *E. coli* strain BL21 (DE3). Rac1 was induced with 1 mM IPTG at a cell density corresponding to $A_{600} = 1.0$ and protein expressed for 6 h at 37 °C. Cells were collected by centrifugation, suspended in 25 mM MES pH 6.0, 50 mM NaCl, 5 mM DTT, 50 μ M GDP, and 1 mM $MgCl_2$ and lysed by passage through a French press. Cell debris was removed by centrifugation and the supernatant adjusted to 30% saturated ammonium sulphate at 4 °C before centrifugation (15 min at 13,000g) of precipitated proteins. Supernatant was further adjusted to 45% saturated ammonium sulphate, and precipitated proteins were centrifuged. The pellet containing Rac1 was resuspended in 20 mM Tris-HCl pH 8.0, 150 mM NaCl, 5 mM DTT, loaded onto a preparative grade S-200 gel exclusion column (26 mM/60 mM), and eluted using the same buffer. Fractions containing Rac1 were pooled, diluted with 20 mM Tris-HCl pH 8.0, 5 mM DTT (buffer D) and loaded onto an 8-ml Source Q column (Pharmacia) equilibrated to buffer D plus 50 mM NaCl before elution with a linear gradient of NaCl.

The Tiam1/Rac1 complex was formed by incubating the two purified proteins at 4 °C in buffer C using about a threefold molar excess of Rac1. The solution containing the two proteins was exchanged into buffer C several times using a stirred cell concentrator (Amicon), and finally concentrated to < 2 ml before loading onto a Superdex 75 gel-exclusion column (Pharmacia; 16 mM/60 mM) equilibrated with buffer C. Fractions containing the complex were pooled and concentrated into crystallization buffer (10 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM DTT and 5% (v/v) glycerol).

Exchange assays were carried out similarly to described methods³⁴. Before addition of Tiam1, 2 μ M of Rac1 (1–177) or bacterially expressed, full-length Cdc42 were incubated with 400 nM methyl-anthranyl GDP (mant-GDP) at 20 °C in a thermostatted cuvette, and fluorescence was measured using a Perkin–Elmer LS-50B ($\lambda_{ex} = 360$ nm, $\lambda_{em} = 440$ nm; slits = 5 per 5 nm). At indicated times, 200 nM Tiam1 or equivalent volume of buffer was added.

Crystallization and data collection

Tiam1/Rac1 crystals were grown by vapour diffusion in sitting drops at 4 °C by adding 5 μ l of well solution (100 mM MES pH 6.0, 18% PEG 3350, 400 mM Li_2SO_4 , 5% glycerol) to 5 μ l of 20 mg ml^{−1} protein complex. Crystals appeared in 2 d and grew as multiple fused plates with individual plates attaining dimensions of 0.1 mm $\times$ 0.4 mm $\times$ 0.4 mm in ~2 weeks. The crystals belong to space group C_2 ($a = 186.9$ Å, $b = 149.3$ Å, $c = 149.2$ Å, $\beta = 121^\circ$). For data collection at 100 K, the solution containing the crystals was adjusted to 100 mM MES pH 6.0, 18% (w/v) PEG 3350, 21% (v/v) glycerol and 80 mM Li_2SO_4 in a stepwise fashion by decreasing the Li_2SO_4 by 40 mM while increasing the glycerol concentration by 2% (v/v). A cryoprotected crystal was then suspended in a rayon loop (Hampton Research) and plunged into liquid nitrogen. A MAD data set was collected using a single crystal on beamline X4A at the National Synchrotron Light Source (Brookhaven National Laboratories). Data were collected at energies corresponding to the minimum of dispersive differences ($\lambda_1 = 12,687.7$ eV), the maximum of anomalous differences ($\lambda_2 = 12,693.9$ eV), and a significant increase in dispersive differences ($\lambda_3 = 12,776.4$ eV) for the selenium atoms present in the crystals. Data were integrated and scaled using DENZO and Scalepack³⁵. Initially, ten selenium atoms were located using averaged anomalous difference coefficients and the peak-searching program SHELXS³⁶. MAD phases were calculated using MLPHARE from the CCP4 suite of programs³⁷, and anomalous difference Fourier maps were used to locate the remaining fourteen selenium atoms. Solvent flattening, histogram matching, and NCS averaging was accomplished using DM¹⁶.

Model building and refinement

The program O³⁸ was used for model building and CNS³⁹ with bulk-solvent correction for automated refinement. The final model contains 16,164 atoms (16,051 protein atoms, 113 solvent atoms) with $R_{cryst} = 26.2\%$ ($R_{free} = 29.3\%$), and has 88.1% of all residues in the core region of the Ramachandran plot with no residues in disallowed regions. The mean temperature factor for all atoms is 59.7 Å² (protein atoms 59.8 Å², solvent atoms 42.9 Å²). The Wilson temperature factor for data in the range $3.14 \text{ Å} > d > 2.84 \text{ Å}$ is 63.6 Å².

Received 23 June; accepted 12 October 2000.

1. Cerione, R. A. & Zheng, Y. The Dbl family of oncogenes. *Curr. Opin. Cell Biol.* **8**, 216–222 (1996).
2. Whitehead, I. P., Campbell, S., Rossman, K. L. & Der, C. J. Dbl family proteins. *Biochim. Biophys. Acta* **1332**, F1–F23 (1997).
3. Van Aelst, L. & D'Souza-Schorey, C. Rho GTPases and signaling networks. *Genes Dev.* **11**, 2295–2322 (1997).
4. Mackay, D. J. & Hall, A. Rho GTPases. *J. Biol. Chem.* **273**, 20685–20688 (1998).
5. Khosravi-Far, R., Campbell, S., Rossman, K. L. & Der, C. J. Increasing complexity of Ras signal transduction: involvement of Rho family proteins. *Adv. Cancer Res.* **72**, 57–107 (1998).
6. Zohn, I. M., Campbell, S. L., Khosravi-Far, R., Rossman, K. L. & Der, C. J. Rho family proteins and Ras transformation: the RHOad less traveled gets congested. *Oncogene* **17**, 1415–1438 (1998).
7. Eva, A. & Aaronson, S. A. Isolation of a new human oncogene from a diffuse B-cell lymphoma. *Nature* **316**, 273–275 (1985).
8. Zheng, Y., Cerione, R. & Bender, A. Control of the yeast bud-site assembly GTPase Cdc42. Catalysis of guanine nucleotide exchange by Cdc24 and stimulation of GTPase activity by Bem3. *J. Biol. Chem.* **269**, 2369–2372 (1994).
9. Hordijk, P. L. *et al.* Inhibition of invasion of epithelial cells by Tiam1–Rac signaling. *Science* **278**, 1464–1466 (1997).

10. Michiels, F., Habets, G. G., Stam, J. C., van der Kammen, R. A. & Collard, J. G. A role for Rac in Tiam1-induced membrane ruffling and invasion. *Nature* **375**, 338–340 (1995).
11. Han, J. *et al.* Lck regulates Vav activation of members of the Rho family of GTPases. *Mol. Cell Biol.* **17**, 1346–1353 (1997).
12. Habets, G. G. *et al.* Identification of an invasion-inducing gene, Tiam-1, that encodes a protein with homology to GDP-GTP exchangers for Rho-like proteins. *Cell* **77**, 537–549 (1994).
13. Habets, G. G., van der Kammen, R. A., Stam, J. C., Michiels, F. & Collard, J. G. Sequence of the human invasion-inducing TIAM1 gene, its conservation in evolution and its expression in tumor cell lines of different tissue origin. *Oncogene* **10**, 1371–1376 (1995).
14. Stam, J. C. *et al.* Targeting of Tiam1 to the plasma membrane requires the cooperative function of the N-terminal pleckstrin homology domain and an adjacent protein interaction domain. *J. Biol. Chem.* **272**, 28447–28454 (1997).
15. Soisson, S. M., Nimnual, A. S., Uy, M., Bar-Sagi, D. & Kuriyan, J. Crystal structure of the Dbl and pleckstrin homology domains from the human Son of sevenless protein. *Cell* **95**, 259–268 (1998).
16. Collaborative Computational Project, Number 4. *Acta Crystallogr. D* **50**, 760–763 (1994).
17. Ferguson, K. M., Lemmon, M. A., Schlessinger, J. & Sigler, P. B. Structure of the high affinity complex of inositol trisphosphate with a phospholipase C pleckstrin homology domain. *Cell* **83**, 1037–1046 (1995).
18. Kubieski, T. J., Chook, Y. M., Parris, W. E., Rozakis-Adcock, M. & Pawson, T. High affinity binding of the pleckstrin homology domain of mSosl to phosphatidylinositol (4,5)-bisphosphate. *J. Biol. Chem.* **272**, 1799–1804 (1997).
19. Anborgh, P. H. *et al.* Ras-specific exchange factor GRF: oligomerization through its Dbl homology domain and calcium-dependent activation of Raf. *Mol. Cell Biol.* **19**, 4611–4622 (1999).
20. Aghazadeh, B. *et al.* Structure and mutagenesis of the Dbl homology domain. *Nature Struct. Biol.* **5**, 1098–1107 (1998).
21. Liu, X. *et al.* NMR structure and mutagenesis of the N-terminal Dbl homology domain of the nucleotide exchange factor Trio. *Cell* **95**, 269–277 (1998).
22. Blomberg, N. & Nilges, M. Functional diversity of PH domains: an exhaustive modelling study. *Fold. Des.* **2**, 343–355 (1997).
23. Rebecchi, M. J. & Scarlata, S. Pleckstrin homology domains: a common fold with diverse functions. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 503–528 (1998).
24. Hirschberg, M., Stockley, R. W., Dodson, G. & Webb, M. R. The crystal structure of human rac1, a member of the rho-family complexed with a GTP analogue. *Nature Struct. Biol.* **4**, 147–152 (1997).
25. Cherfils, J. & Chardin, P. GEFs: structural basis for their activation of small GTP-binding proteins. *Trends Biochem. Sci.* **24**, 306–311 (1999).
26. Steven, R. *et al.* UNC-73 activates the Rac GTPase and is required for cell and growth cone migrations in *C. elegans*. *Cell* **92**, 785–795 (1998).
27. Kawashima, T., Berthet-Colominas, C., Wulff, M., Cusack, S. & Leberman, R. The structure of the *Escherichia coli* EF-Tu:EF-Ts complex at 2.5 Å resolution. *Nature* **379**, 511–518 (1996).
28. Boriack-Jodini, P. A., Margalit, S. M., Bar-Sagi, D. & Kuriyan, J. The structural basis of the activation of Ras by Sos. *Nature* **394**, 337–343 (1998).
29. Goldberg, J. Structural basis for activation of ARF GTPase: mechanisms of guanine nucleotide exchange and GTP-myristoyl switching. *Cell* **95**, 237–248 (1998).
30. Shimizu, T. *et al.* An open conformation of switch I revealed by the crystal structure of a Mg^{2+} -free form of RHOA complexed with GDP. Implications for the GDP/GTP exchange mechanism. *J. Biol. Chem.* **275**, 18311–18317 (2000).
31. Leonard, D. *et al.* The identification and characterization of a GDP-dissociation inhibitor (GDI) for the CDC42Hs protein. *J. Biol. Chem.* **267**, 22860–22868 (1992).
32. Hoffman, G. R., Nassar, N. & Cerione, R. A. Structure of the Rho family GTP-binding protein Cdc42 in complex with the multifunctional regulator RhoGDI. *Cell* **100**, 345–356 (2000).
33. Han, J. *et al.* Role of substrates and products of PI 3-kinase in regulating activation of Rac-related guanosine triphosphatases by Vav. *Science* **279**, 558–560 (1998).
34. Lenzen, C., Cool, R. H. & Wittinghofer, A. Analysis of intrinsic and Cdc25-stimulated guanine nucleotide exchange of p21ras-nucleotide complexes by fluorescence measurements. *Methods Enzymol.* **255**, 95–109 (1995).
35. Otwinowski, J. *Data Collection and Processing* (eds Sawyer, L., Isaacs, N. & Bailey, S.) (Daresbury Laboratory, Warrington, UK, 1993).
36. Sheldrick, G. M. *SHELXS86 – Program for Crystal Structural Solution* (Univ. Göttingen, Göttingen, Germany, 1986).
37. Otwinowski, J. (ed.) *Maximum Likelihood Refinement of Heavy Atom Parameters* (Daresbury, Warrington, UK, 1991).
38. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
39. Brünger, A. T. *et al.* A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
40. Kraulis, P. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).
41. Merritt, E. A. & Murphy, M. E. P. Raster3D Version 2.0. A program for photorealistic molecular graphics. *Acta Crystallogr. D* **50**, 869–873 (1994).
42. Carson, M. Ribbons. *Methods Enzymol.* **277**, 493–505 (1997).
43. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and associations: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).

Supplementary information is available on *Nature's* World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We would like to thank M. Pham and S. Foster for technical assistance; J. Snyder for assistance with data collection; L. Betts, H. Ke, C. Der, D. Siderovski, B. Worthyake, A. Singer, and C. Winkelman for critical reading of this manuscript; C. Ogata, and R. Abramowitz of beamline X4A, and the staff of NSLS for help with synchrotron data collection; D.K.W. is supported by a grant from the American Cancer Society, J.S. acknowledges support from the NIH and the Pew Charitable Trusts.

Correspondence and request for materials should be addressed to J.S. The atomic coordinates have been deposited at the RCSB under code 1FOE.

An optical counterpart to the anomalous X-ray pulsar 4U0142+61

F. Hulleman*, M. H. van Kerkwijk* & S. R. Kulkarni†

*Astronomical Institute, Utrecht University, PO Box 80000, 3508 TA Utrecht, The Netherlands

†Palomar Observatory 105-24, California Institute of Technology, Pasadena, California 91125, USA

The energy source of the anomalous X-ray pulsars¹ (AXPs) is not understood, hence their designation as anomalous. Unlike binary X-ray pulsars, no companions are seen, so the energy cannot be supplied by accretion of matter from a companion star. The loss of rotational energy, which powers radio pulsars, is insufficient to power AXPs. Two models are generally considered: accretion from a large disk left over from the birth process^{2,3}, or decay of a very strong magnetic field (10^{15} G) associated with a 'magnetar'⁴. The lack of counterparts at other wavelengths has hampered progress in our understanding of these objects. Here we report deep optical observations of the field around 4U0142+61, which is the brightest

AXP in X-rays. The source has no associated supernova remnant, which, together with its spin-down timescale of $\sim 10^5$ yr (ref. 5), suggests that it may be relatively old. We find an object with peculiar optical colours at the position of the X-ray source, and argue that it is the optical counterpart. The optical emission is too faint to admit the presence of a large accretion disk, but may be consistent with magnetospheric emission from a magnetar.

The field around 4U0142+61 was imaged in the V, R and I bands on 31 October 1994 using the low-resolution imaging spectrometer⁶ on the 10-m Keck I telescope on Mauna Kea, and in R and I on 6 September 1999 on the Keck II telescope. In 1994, the sky was not photometric and the seeing mediocre at 1 arcsec, but in 1999 the sky was photometric and the seeing was 0.6 arcsec. Photometric calibration was achieved with images obtained under photometric conditions on 23 July 2000 at the 60-inch telescope on Palomar Mountain.

Within the error circles of 4U0142+61 (Fig. 1), we find one faint stellar object with unusual colours (Fig. 2). To understand its nature, and its relation to 4U0142+61, we need to constrain the run of reddening with distance along the line of sight. We use two open clusters in this region of sky⁷—NGC654 (at an angular offset of $17'$, distance of 2.7 ± 0.4 kpc and reddening $A_V = 2.3$ – 3.4) and NGC663 ($30'$, 2.8 ± 0.4 kpc, 2.2 – 3.1)—to infer that $A_V \approx 2.7$ at

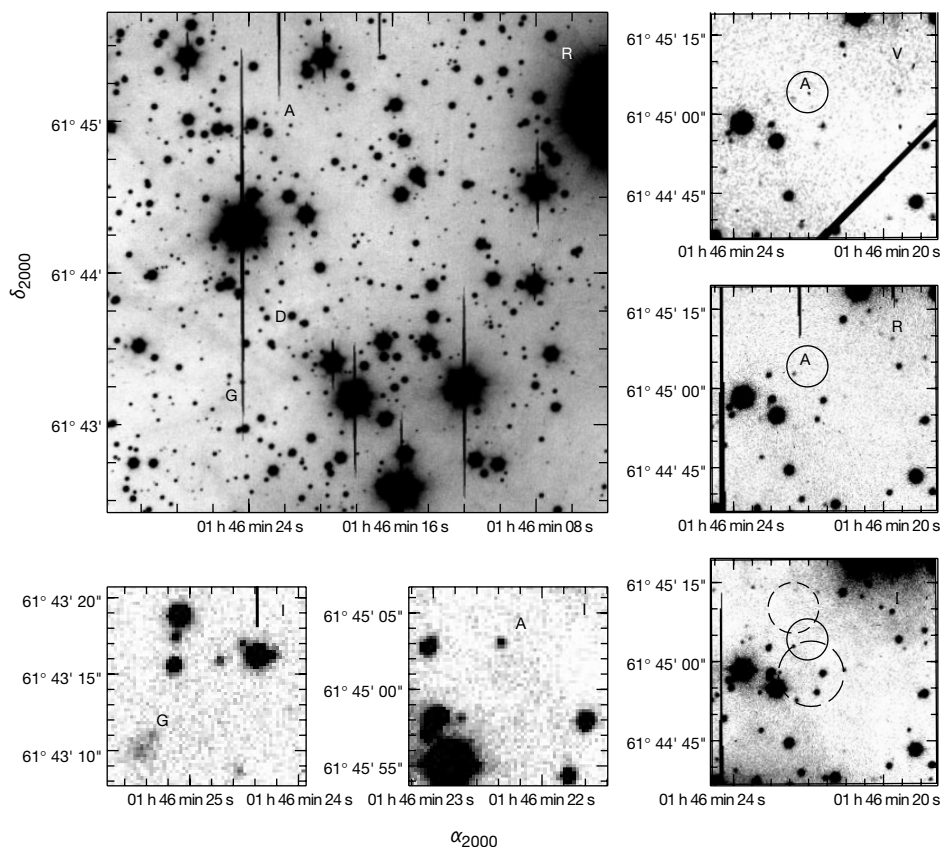


Figure 1 Keck images around the X-ray position of 4U0142+61. The large panel shows an overview of the R-band image. Our candidate, A, as well as two other blue objects, D and G, are indicated with labels just to the right of their images. The small bottom panels are close-ups of the I-band images of A and G, which show that G is extended and is probably a galaxy, while A is not. The panels on the right show the V, R, and I images around star A. The error circle indicates the X-ray position derived from analysis of the Einstein High Resolution Imager (HRI) observations ($\alpha_{J2000} = 01\text{ h }46\text{ min }22.03\text{ s}$, $\delta_{J2000} = +61^\circ 45' 04.3''$, 95% confidence level radius of 3.9 arcsec; ref. 22). The Einstein positions have proven to be highly reliable, but for comparison we display in the I-band panel also the error circles derived from observations with the EXOSAT LEIT (end figures 22.40 s and 10.2"; 4.8 arcsec radius; ref. 22) and ROSAT PSPC (21.93 s,

44' 57.7"; 6 arcsec radius). The ROSAT position of 4U0142+61 was derived from archival data, using a boresight correction inferred from two other X-ray sources in the field, for which the optical counterparts are known^{23,24} (LS I +61° 235 and LP 80-77; our technique is described in ref. 14). The astrometry was done as in ref. 14, by measuring 107 stars from the USNO-A2.0 catalogue²⁵ on a short R-band exposure (root-mean-square residuals of 0.26 arcsec in each coordinate, consistent with the USNO-A2.0 measurement errors). From the I-band image, we infer a position of star A of $\alpha_{J2000} = 01\text{ h }46\text{ min }22.41\text{ s}$, $\delta_{J2000} = +61^\circ 45' 03.2''$. We estimate that this position is on the international celestial reference frame to about 0.2 arcsec. The other images give consistent positions. Our 2σ upper limit to the proper motion is $0.03\text{ arcsec yr}^{-1}$.

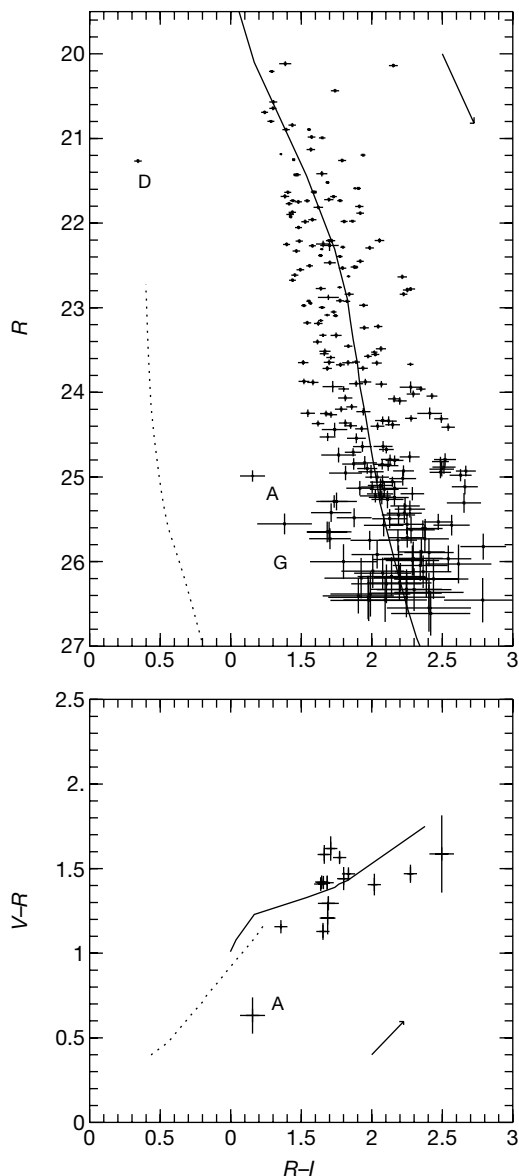


Figure 2 Colour-magnitude and colour-colour diagrams for stars near the position of 4U0142+61. For comparison, the magnitudes and colours expected for stars²⁶ at the distance, reddening, and age of the nearby open cluster NGC654 (2.7 ± 0.4 kpc, 2.7 mag, 50 Myr; ref. 7; solid line) are shown, as well as a cooling track for a $0.6 M_{\odot}$ white dwarf²⁷ at the same reddening and distance (dashed line). The arrows indicate the effect of increasing the reddening by $\Delta A_V = 1$. All optical images were corrected for sensitivity variations using dome flats, and a small correction for nonlinearity of the detector was applied to the 1999 observations. For the 1999 I-band observations, a 'fringe' frame was constructed by taking the second-lowest value for each pixel from seven dithered images, which was subtracted for the individual images to remove interference patterns. Instrumental magnitudes were measured using the DAOPHOT II package²⁸ on the 1994 V band and the 1999 R- and I-band images, using only those regions not affected by scattered light from bright, overexposed stars. The 1994 V-band images were taken at a different pointing and rotation angle, which caused different parts on the sky to be affected by bleed trails, and so on. Therefore, we obtained V-band measurements only for stars in the central part displayed in Fig. 1. The photometric calibration was done relative to the 60-inch data, which were calibrated in turn using exposures of the standard fields PG1657-042 and NGC7790 (ref. 29). The R and I calibration was verified using Keck images of the standard-field PG0231+051 (ref. 29). We estimate the uncertainties in the zero points of the magnitudes and colours to be about 0.03 mag. For star A, we find $R = 24.99 \pm 0.07$, $V - R = 0.63 \pm 0.11$, and $R - I = 1.15 \pm 0.09$. By comparison with the 1994 images, we find that its brightness was constant to within 0.2 mag (2σ) in the R band.

$d = 2.7$ kpc. Furthermore, we use the bluest galaxy in our field (object G), with $R = 25.6$, $R - I = 1.3$. Because the intrinsic colour⁸ is unlikely to be bluer than $(R - I)_0 = 0.3$, the total Galactic reddening along this line of sight is constrained by $E_{R-I} \leq 1.1$, or equivalently $A_V \leq 5$.

From Fig. 2, we see that most stars have brightnesses and colours consistent with those expected for low-mass stars at distances similar to the open clusters. Star A, however, is either too dim or too blue. It might be bluer because it is nearer and therefore less reddened. But then it would be too dim to be a main-sequence star and too red to be a white dwarf. The latter can be seen by comparison to star D, which is probably a foreground white dwarf (Fig. 2).

We proceed to estimate the temperature T and radius R of star A on the assumption that the object is located beyond the open clusters (that is, $d > 2.7$ kpc; $2.7 \leq A_V \leq 5$); below we use d_5 , the distance normalized to 5 kpc. Bearing in mind the intrinsic variability in the shape of the reddening curve⁹, we find a marginal fit for $T \approx 6,000$ K and $R \approx 0.11 d_5 R_{\odot}$ ($A_V = 2.7$), where $R_{\odot}$ is the Sun's

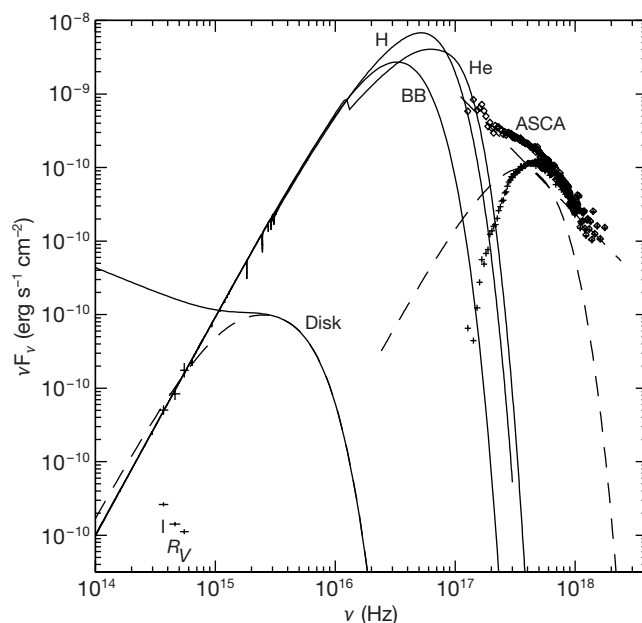


Figure 3 Energy distribution for 4U0142+61. At low frequencies (10^{14} – 10^{15} Hz), the points marked V, R and I indicate the observed V-, R-, and I-band fluxes. The vertical error bars reflect the uncertainties, while the horizontal ones indicate the filter bandwidths. The set of points above the measurements indicate dereddened fluxes for $A_V = 5.4$, as inferred from the X-ray column density^{10,11}. The errors include a 3% uncertainty in the reddening correction⁹. At high frequencies (10^{17} – 10^{18} Hz), the crosses show the incident X-ray spectrum as inferred from ASCA measurements¹⁰. The diamonds show the spectrum after correction for interstellar absorption, and the two thick dashed curves show the two components used in the fit¹⁰: a power law of the form $F_{\nu} = 103(h\nu/1 \text{ keV})^{-2.67} \mu\text{Jy}$ and a black body with $T = 4.4 \times 10^6$ K and $R = 12 d_5$ km. The latter component is extrapolated to lower frequencies to show that it cannot reproduce the optical fluxes. Drawn in thin solid lines are models for the optical emission. The curve marked 'Disk' is for an accretion disk that has inner radius equal to the co-rotation radius $R_{\text{co}} = (GMP^2/4\pi^2)^{1/3} = 0.010 R_{\odot}$ (for $M = 1.4 M_{\odot}$, $P = 8.7$ s), outer radius of 10^{14} cm, and inclination of 60 degrees. Both irradiation and viscous heating are taken into account¹³; the former dominates in the optical and the latter causes the bump in the ultraviolet. The calculation is for $d = 5$ kpc, but since the disk luminosity scales almost linearly with the X-ray luminosity, the result is not sensitive to distance. The optical fluxes expected in the accretion model are greatly in excess of those observed. Only for a truncated disk with a small outer disk radius, $r_{\text{out}} \leq 0.1 d_5^{0.11} (f/0.25)^{-2/11} R_{\odot}$, can the observed fluxes be reproduced, as is shown by the dashed curve. The thin solid curves marked BB, H, and He are black body, pure hydrogen and pure helium white dwarf model atmospheres for $T = 4 \times 10^5$ K. The normalization to the optical data implies $R = 0.011 d_5$, $0.017 d_5$ and $0.015 d_5 R_{\odot}$, respectively. None of the spectra can reproduce the X-ray emission.

radius; and increasingly better fits for $T \approx 10^4$ K and $R \approx 0.08d_5R_\odot$ ($A_V \approx 4$) up to $T \geq 10^5$ K and $R \approx 0.022d_5(T/10^5 \text{ K})^{-1/2}R_\odot$ ($A_V \approx 5.4$). We note that for $T \geq 10^5$ K the optical bands fall in the Rayleigh–Jeans tail and the colours do not vary with T any more. Because temperature and reddening compensate each other, the above scaling in the Rayleigh–Jeans limit actually holds to within 10% for any $T \geq 6,000$ K.

For all temperatures, the radii are too small for star A to be a normal star within our Galaxy. The remaining possibility is that star A is a reddened but hot ($T \approx 5 \times 10^5 d_5^2$ K) white dwarf. This is extremely unlikely unless it is actually 4U0142+61 itself, as we discuss below.

From the above, we conclude that in all likelihood star A is the optical counterpart to 4U0142+61. Below, we discuss our observations in the framework of the models that have been proposed for AXPs. The optical extinction A_V to 4U0142+61 is between 2.7 and 5.4 mag, as estimated from the X-ray dust-scattering halo and absorption column density, respectively^{10,11} (where the latter is probably the more reliable¹⁰); therefore, the source must be at a distance exceeding about 2.7 kpc.

We first consider the possibility that AXPs are isolated neutron stars accreting matter from a disk, presumably composed of supernova debris^{2,3}. In this model, the optical emission arises in the accretion disk, mostly owing to reprocessing of the X-ray irradiation¹² (Fig. 3). The isotropic X-ray luminosity of 4U0142+61 is $L_X = 4\pi d^2 f_X^{\text{unabs}} = 1.8 \times 10^{36} d_5^2 \text{ erg s}^{-1}$, where f_X^{unabs} is the flux in the energy range 0.5–10 keV after correction for interstellar absorption (see Table 1). We can estimate the run of temperature with disk radius¹³ r by $T(r) \approx 5,000(f/0.25)^{2/7} d_5^{4/7} (r/R_\odot)^{-3/7}$ K. Here, the factor f parametrizes uncertainties in vertical disk structure and the fraction of impinging X-ray emission reflected by the disk.

The flux at a given frequency and for given inclination is obtained by integrating the emission from the accretion disk (assumed to be optically thick) from an inner radius, r_{in} , to an outer radius, r_{out} ; see ref. 12 and Fig. 3. For a simple estimate, we use the fact that the optical flux will arise predominantly at those radii at which the emission peaks in the optical band, that is, where $T \approx 5,000$ K. For 4U0142+61, this would be at $r \approx 1(f/0.25)^{2/3} d_5^{4/3} R_\odot$, which is much larger than the limit of $R \approx 0.11d_5R_\odot$ derived above, and leads to a brightness much larger than observed (see Fig. 3). Thus, this model is excluded.

The optical emission from the disk can be reduced by appealing to a larger r_{in} or a smaller r_{out} . The former, suggested¹² in the context of earlier limits¹⁴, would result in extremely red emission (because of the absence of the hotter inner region). This is incompatible with the observations.

A reduced r_{out} would follow naturally in a model less often considered, in which the AXP are compact binaries¹⁵. We find that our observations require $r_{\text{out}} \approx 0.05d_5^{10/11} (f/0.25)^{-2/11} R_\odot$; see Fig. 3.

This is very small, but not unprecedented: the X-ray binary with the tightest known orbit¹⁶, 4U1820–30 (an 11-min orbital period) has a similarly small radius and similarly high ratio of X-ray to optical flux; see Table 1. However, the X-ray spectrum of 4U0142+61 is rather different from that of 4U1820–30 and other accreting sources. Furthermore, such compact binaries are not expected to be associated preferentially with supernova remnants. Thus, we consider this model unlikely.

As mentioned, the optical data are consistent with a hot white dwarf. This would be expected in another model rarely considered, in which AXPs are massive and magnetized (10^8 G) white dwarfs rotating on the verge of break-up, presumably formed in the merger of two ordinary, approximately $0.6 M_\odot$ white dwarfs¹⁷. In this model, the ultimate source of energy, as in radio pulsars, is rotation. For 4U0142+61, the inferred rotational energy loss, $-\dot{E}_{\text{rot}} = 4\pi^2 I \dot{P}/P^3 \approx 10^{37} \text{ erg s}^{-1}$, is clearly sufficient to account for the X-ray luminosity; here $I = kMR^2 \approx 10^{50} \text{ g cm}^2$ is the moment of inertia appropriate for a hot white dwarf with mass $M \approx 1.3 M_\odot$, where $M_\odot$ is the Sun's mass, radius $R \approx 0.007 R_\odot$ and gyration constant $k \approx 0.14$. P is the spin period and $\dot{P}$ its derivative. If some fraction β of $\dot{E}_{\text{rot}}$ goes into heating, as might result if the white dwarf rotated differentially, we expect a surface luminosity $L = 4\pi R^2 \sigma T^4 = \beta \dot{E}_{\text{rot}}$, (where σ is the Stefan–Boltzmann constant) and thus a surface temperature $T \approx 4 \times 10^5 (\beta/0.5)^{1/4} (k/0.14)^{1/4} (M/1.3 M_\odot)^{1/4}$ K (we note that T does not depend on R). This is sufficient to produce the optical emission for a source at $d_5 = 0.6(R/0.007 R_\odot)$ (see Fig. 3). We note that a possible descendant may already have been identified¹⁸: the $1.3 M_\odot$ white dwarf RE J0317–853 with $P = 725$ s, $B \approx 5 \times 10^8$ G, and a cooling age of a few 10^8 yr. In this model, however, what mechanism may cause the X-ray spectrum is unclear, the thermal emission being far too soft (see Fig. 3). Furthermore, the association of other AXPs with supernova remnants again seems puzzling.

This leaves us with the magnetar model⁴, in which both the power-law component of the X-ray emission and the optical emission would be of magnetospheric origin. Unfortunately, there are no detailed models. Radio pulsars with X-ray and optical detections have rather smaller f_X/f_{opt} (see Table 1), but their X-ray spectra are unlike those of AXPs. From Fig. 3, it appears that the simplest spectral energy distribution would be obtained if the extinction were near the high end ($A_V \geq 5$), so that the emission could peak at a few 10^{16} Hz (~ 100 eV) and be like a power law $F_\nu \propto \nu^\alpha$, where F_ν is the flux per unit frequency at frequency ν , with index $\alpha \approx 2$ (or even 2.5) in the optical. This could arise if the magnetospheric emission is self-absorbed at optical frequencies. A possible problem for the magnetar model is the lack of X-ray emission from the one radio pulsar with similar P and $\dot{P}$ (ref. 19). However, it may well be that P and $\dot{P}$ are not reliable measures of the

Table 1 Comparison with other high-energy objects

Object	Type	P_{spin} (s)	A_R	R	$(V-R)_0$	$(R-I)_0$	f_X^{unabs} ($\text{erg s}^{-1} \text{ cm}^{-2}$)	$f_X^{\text{unabs}}/f_R^{\text{unabs}}$	d (kpc)
4U0142+61	AXP	8.69	4.4	24.98	–0.33	0.01	6.0×10^{-10}	7.1×10^3	>2.7
			2.1		0.22	0.64		5.9×10^4	
1E2259+58.6	AXP	6.97	3.8	>25.7			1.8×10^{-10}	$\geq 7.2 \times 10^3$	≥ 5
Crab	PSR	0.033	1.3	16.21	–0.06	0.35	4.0×10^{-9}	2.5×10^2	2
Vela	PSR	0.089	0.3	23.93	–0.35		7.5×10^{-12}	1.4×10^3	0.4:
PSR B0656+14	PSR	0.38	0.1	24.52	0.36	0.69	1.5×10^{-12}	6.1×10^2	0.5:
Geminga	PSR	0.24	0.1	25.4	–0.20		3.9×10^{-12}	3.6×10^3	0.2:
4U1626–67	XRB	7.66	0.2	18.68	–0.06	0.82:	$(2-8) \times 10^{-10}$	$(3-14) \times 10^2$	8::
4U1820–30	XRB		0.7	18.87:			$(8-16) \times 10^{-9}$	$(1-2) \times 10^4$	8
RX J0720.4–3125	INS	8.39	0.1	26.9			2.5×10^{-12}	9.1×10^3	<0.4

Type of object: AXP: anomalous X-ray pulsar; PSR: radio pulsar; XRB: X-ray binary; INS: isolated neutron star. Although RX J0720.4–3125 has a similar pulse period and X-ray to optical flux ratio, its X-ray spectrum is very different from those of the AXPs. The observational data have been collected from the literature. A full set of references can be found on <http://www.astro.uu.nu/~hulleman/u0142>. The unabsorbed X-ray fluxes are for the energy range of 0.5–10 keV. They have been inferred for the best-fit spectral model with the help of the PIMMS software package. For 4U1626–67, constant pulsed flux is assumed in the estimate of $(R-I)_0$, and for 4U1820–30, we assumed $V-R=0$ to estimate R . For the conversion to fluxes, we used the Bessel zero points²⁰, where $V=0$ corresponds to 3,600 Jy, $R=0$ to 3,060 Jy, and $I=0$ to 2,420 Jy. For reddening corrections, we used $(A_V, A_R, A_I) = (1.015, 0.819, 0.594)A_V$, where A_V, A_R , and A_I are reddenings in the Landolt V, R, and I filters, and A_V is the reddening in the Johnson V band; see Appendix C in ref. 21 for a discussion. For reference, the final column lists estimated distances; the colons indicate the uncertainty, zero colons indicating an uncertainty factor of up to 1.5, one colon a factor of two, and two colons a factor of three.

strength of magnetic fields for AXPs and that other properties determine whether a source is an AXP.

In spite of the current uncertainties, the optical identification offers us new insights into these enigmatic objects and motivates new observations and searches for optical pulsations. □

Received 11 August; accepted 31 October 2000.

1. Mereghetti, S. in *Proc. NATO ASI School "The Neutron Star—Black Hole Connection"* (eds Kouveliotou, C., van Paradijs, J. & Ventura, J.) (Kluwer Academic Publishers, Dordrecht, 1999); also preprint astro-ph/9911252 at (xxx.lanl.gov) (1999).
2. Marsden, D., Lingenfelter, R. E., Rothschild, R. E. & Higdon, J. C. Nature vs. nurture: The origin of soft gamma-ray repeaters and anomalous X-ray pulsars. *Astrophys. J.* (submitted); also preprint astro-ph/9912207 at (xxx.lanl.gov) (1999).
3. Chatterjee, P., Hernquist, L. & Narayan, R. An accretion model for anomalous X-ray pulsars. *Astrophys. J.* **534**, 373–379 (2000).
4. Thompson, C. & Duncan, R. C. The soft gamma repeaters as very strongly magnetized neutron stars. II. Quiescent neutrino, X-ray, and alfvén wave emission. *Astrophys. J.* **473**, 322–342 (1996).
5. Wilson, C. A., Dieters, S., Finger, M. H., Scott, D. M. & van Paradijs, J. Rossi X-ray timing explorer observations of the anomalous pulsar 4U0142+61. *Astrophys. J.* **513**, 464–470 (1999).
6. Oke, J. B. *et al.* The Keck low-resolution imaging spectrometer. *Publ. Astron. Soc. Pacif.* **107**, 375–385 (1995).
7. Phelps, R. L. & Janes, K. A. Young open clusters as probes of the star formation process. I: An atlas of open cluster photometry. *Astrophys. J.* **90**, 31–82 (1994).
8. Fukugita, M., Shimasaku, K. & Ichikawa, T. Galaxy colors in various photometric band systems. *Publ. Astron. Soc. Pacif.* **107**, 945–958 (1995).
9. Cardelli, J. A., Clayton, G. C. & Mathis, J. S. The relationship between infrared, optical, and ultraviolet extinction. *Astrophys. J.* **345**, 245–256 (1989).
10. White, N. E., Angelini, L., Ebisawa, K., Tanaka, Y. & Ghosh, P. The spectrum of the 8.7s X-ray pulsar 4U0142+61. *Astrophys. J.* **463**, L83–L86 (1996).
11. Predehl, P. & Schmitt, J. H. M. X-raying the interstellar medium: ROSAT observations of dust scattering halos. *Astron. Astrophys.* **293**, 889–905 (1995).
12. Perna, R., Hernquist, L. & Narayan, R. Emission spectra of fall back disks around young neutron stars. *Astrophys. J.* **541**, 344–350 (2000).
13. Vrtillek, S. D. *et al.* Observations of Cygnus X-2 with IUE—Ultraviolet results from a multiwavelength campaign. *Astron. Astrophys.* **235**, 162–173 (1990).
14. Hulleman, F., van Kerkwijk, M. H., Verbunt, F. W. M. & Kulkarni, S. R. A deep search for the optical counterpart to the anomalous X-ray pulsar 1E 2259+58.6. *Astron. Astrophys.* **358**, 605–611 (2000).
15. Mereghetti, S. & Stella, L. The very low mass X-ray binary pulsars: A new class of sources? *Astrophys. J.* **422**, L17–L20 (1995).
16. Stella, L., White, N. E. & Priedhorsky, W. The discovery of a 685 second orbital period from the X-ray source 4U 182030 in the globular cluster NGC 6624. *Astrophys. J.* **312**, L17–L21 (1987).
17. Paczynski, B. X-ray pulsar 1E 2259+586—A merged white dwarf with a 7 second rotation period? *Astrophys. J.* **365**, L9–L12 (1990).
18. Ferrario, L., Vennes, S., Wickramasinghe, D. T., Bailey, J. A. & Christian, D. J. EUVE J0317–855 A rapidly rotating, high-field magnetic white dwarf. *Mon. Not. R. Astron. Soc.* **292**, 205–217 (1997).
19. Pivovarov, M. J., Kaspi, V. M. & Camilo, F. X-ray observations of the high magnetic field radio pulsar PSR J1814–1744. *Astrophys. J.* **535**, 379–384 (2000).
20. Bessell, M. S. in *Proc. IAU Coll. 136, Stellar Photometry—Current Techniques and Future Developments* (eds Butler, C. J. & Elliott, I.) 22–39 (Cambridge Univ. Press, Cambridge, 1992).
21. Schlegel, D. J., Finkbeiner, D. P. & Davies, M. Maps of dust infrared emission for use in estimation of reddening and cosmic microwave background radiation foregrounds. *Astrophys. J.* **500**, 525–553 (1998).
22. White, N. E. *et al.* A 25 min modulation from the vicinity of the unusually soft X-ray source X0142+614. *Mon. Not. R. Astron. Soc.* **226**, 645–654 (1987).
23. Motch, C. *et al.* A ROSAT glance at the galactic plane. *Astron. Astrophys.* **246**, L24–L27 (1991).
24. Van den Berg, M. & Verbunt, F. Spectroscopic confirmation of the optical identification of X-ray sources used to determine accurate positions for the anomalous X-ray pulsars 1E 2259+58.6 and 4U 0142+61. *Astron. Astrophys.* (submitted).
25. Monet, D. *et al.* USNO-A V2.0, A Catalog of Astrometric Standards (US Naval Observatory Flagstaff Station (USNOFS) and Universities Space Research Association (USRA) stationed at USNOFS, 2000).
26. Baraffe, I., Chabrier, G., Allard, F. & Hauschildt, P. H. Evolutionary models for solar metallicity low-mass stars: mass-magnitude relationships and color-magnitude diagrams. *Astron. Astrophys.* **337**, 403–412 (1998).
27. Bergeron, P., Wesemael, F. & Beauchamp, A. Photometric calibration of hydrogen- and helium-rich white dwarf models. *Publ. Astron. Soc. Pacif.* **107**, 1047–1054 (1995).
28. Stetson, P. B. DAOPHOT—A computer program for crowded-field stellar photometry. *Publ. Astron. Soc. Pacif.* **99**, 191–222 (1987).
29. Stetson, P. B. Homogeneous photometry for star clusters and resolved galaxies. II. Photometric standard stars. *Publ. Astron. Soc. Pacif.* **112**, 925–931 (2000).

Acknowledgements

We thank D. Kaplan for undertaking the observations at the 60-inch telescope, and acknowledge helpful discussions with B. Paul and F. Verbunt. Model atmospheres for hot white dwarfs were kindly provided by J. Heise. The observations were obtained at the W. M. Keck Observatory on Mauna Kea, Hawaii, which is operated by the California Association for Research in Astronomy. This research made use of the SIMBAD data base. M.H.v.K. is supported, in part, by grants from the National Science Foundation and the National Aeronautics and Space Administration.

Correspondence and requests for materials should be addressed to F.H. (e-mail: f.hulleman@astro.uu.nl).

Direct observation of growth and collapse of a Bose–Einstein condensate with attractive interactions

Jordan M. Gerton*, Dmitry Strekalov†, Ionut Prodan* & Randall G. Hulet*

* Physics and Astronomy Department and Rice Quantum Institute, MS 61, Rice University, Houston, Texas 77251, USA

† Present address: Jet Propulsion Laboratory, MS 300-123, Pasadena, California 91109, USA.

Quantum theory predicts that Bose–Einstein condensation of a spatially homogeneous gas with attractive interactions is precluded by a conventional phase transition into either a liquid or solid¹. When confined to a trap, however, such a condensate can form², provided that its occupation number does not exceed a limiting value^{3,4}. The stability limit is determined by a balance between the self-attractive forces and a repulsion that arises from position–momentum uncertainty under conditions of spatial confinement. Near the stability limit, self-attraction can overwhelm the repulsion, causing the condensate to collapse^{5–8}. Growth of the condensate is therefore punctuated by intermittent collapses^{9,10} that are triggered by either macroscopic quantum tunnelling or thermal fluctuation. Previous observations of growth and collapse dynamics have been hampered by the stochastic nature of these mechanisms. Here we report direct observations of the growth and subsequent collapse of a ⁷Li condensate with attractive interactions, using phase-contrast imaging. The success of the measurement lies in our ability to reduce the stochasticity in the dynamics by controlling the initial number of condensate atoms using a two-photon transition to a diatomic molecular state.

Condensate growth is initiated by cooling the gas below the critical temperature T_c for Bose–Einstein condensation (BEC). For

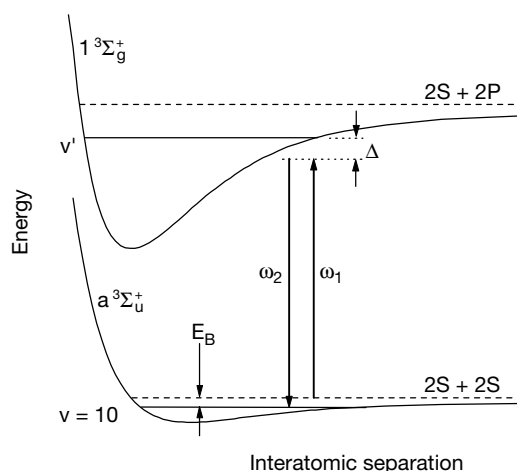


Figure 1 Two-photon molecular association technique. Two atoms in the electronic ground state ($2S + 2S$) interact along the $a^3\Sigma_u^+$ diatomic molecular potential at the dissociation limit indicated by the dashed line. They may be associated into the least-bound vibrational level, $v = 10$, when the relative frequency between two drive lasers, $\omega_1 - \omega_2$, is equal to E_B , where $E_B \approx 12.47$ GHz is the binding energy of the $v = 10$ level. Each laser is tuned near resonance with a vibrational level $v' = 72$ of the $1^3\Sigma_g^+$ excited state potential, which correlates to $2S + 2P$ atomic states. The intensity of each laser beam is between 0.5 and 2.5 W cm⁻² and the intermediate state detuning $\Delta \approx -110$ MHz.

attractive interactions, the number of condensate atoms N_0 grows until the condensate collapses. During the collapse, the density rises, giving a sharp increase in the rate of collisions, both elastic and inelastic. These collisions cause atoms to be ejected from the condensate in an energetic explosion. The physics that determines both the stability condition and the dynamical process of collapse of the condensate bears some similarity to that of a star going supernova¹¹, even though the time, length, and energy scales for these two phenomena are very different. In the stellar case, the stability criterion is provided by a balance between the pressure due to the quantum degeneracy of electrons that make up the star and gravitational attraction. If the mass of the star exceeds the stability limit¹², the star collapses, releasing nuclear energy and triggering a violent explosion. In contrast to the stellar case, the condensate regrows after a collapse as it is fed by collisions between thermal atoms in the gas, and a series of sawtooth-like cycles of growth and collapse will continue until the gas reaches thermal equilibrium^{9,10}. We previously attained evidence for this nonequilibrium dynamical behaviour in ^7Li by measuring the distribution of N_0 at selected times after a fast quench of the gas¹³. N_0 was found to be distributed between small numbers, $N_0 \approx 100$, and the maximum number of $\sim 1,250$ atoms, in agreement with the growth and collapse model.

The process of making reliable measurements of such small values of N_0 destroys the condensate, preventing an observation of condensate dynamics in real time. Although the phase-contrast imaging technique employed here has been used for (nearly) nondestructive measurements of large condensates¹⁴, a few incoherent photons are always scattered, and these heat the gas. As the phase-contrast signal is proportional to the number of scattered photons, achieving sufficient sensitivity to small condensates, such as those studied here, results in excessive heating and destruction of the condensate¹⁵. Here, direct observation of the dynamics is made possible by dumping the condensate, while only slightly affecting the thermal atoms. This synchronizes the growth and collapse cycles for different experiments at a particular point in time. The subsequent growth/collapse dynamics are then obtained by repeating the experiment and measuring N_0 at different delays following the dump pulse. In a recent experiment with essentially pure condensates of ^{85}Rb atoms, a magnetically tuned Feshbach resonance was used to suddenly switch the interactions from repulsive to attractive, thereby inducing a collapse at a specified time¹⁶. In that experiment, condensates are produced with N_0 far greater than the stability limit, and consequently with high initial energy. In

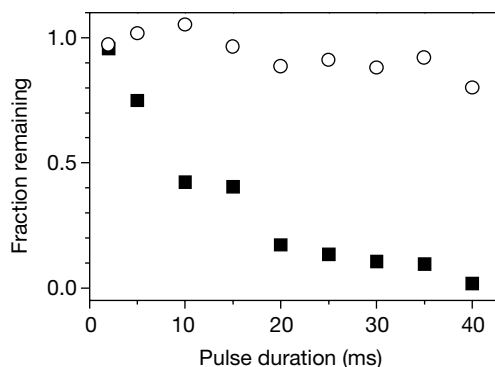


Figure 2 Energy selectivity of the two-photon light pulse. The fraction of condensate atoms (solid squares) and total atoms (open circles) which remain trapped immediately after a light pulse is plotted as a function of the light pulse duration. Each data point is an average of five separate experiments normalized to measurements taken without the light pulse. For these measurements, $N = (7 \pm 1) \times 10^4$ atoms. Achieving the necessary energy selectivity required the two diode lasers used to drive the two-photon transition to be phase-locked. When locked, the relative frequency spectrum of the two lasers was measured to be less than 1 Hz in width.

contrast, for the present experiment, the collapse occurs with N_0 below the maximum number by macroscopic quantum tunnelling or thermal fluctuation⁷⁻⁹. Furthermore, the condensate coexists with a gas of thermal atoms, allowing the kinetics to be probed.

The apparatus used to produce BEC of ^7Li has been described previously¹⁵. Permanent magnets establish an Ioffe-Pritchard type trap with a nearly spherically symmetric, harmonic trapping potential. Approximately 2.5×10^8 atoms in the $F = 2$, $m_F = 2$ hyperfine sublevel of ^7Li are directly loaded into the trap from a laser-slowed atomic beam. Following loading, the atoms are evaporatively cooled to ~ 400 nK with $\sim 4 \times 10^5$ atoms using a microwave field to selectively spin-flip and remove the most energetic atoms. Under these conditions, the gas is quantum degenerate. The microwave frequency is maintained at the end frequency for two seconds after evaporation, and is subsequently lowered during a quench pulse 100 ms in duration that removes all but the coldest $\sim 10^5$ atoms. This leaves the gas far from thermal equilibrium, and if left to freely evolve, the condensate will alternately grow and collapse many times over a period of ~ 30 s (ref. 13).

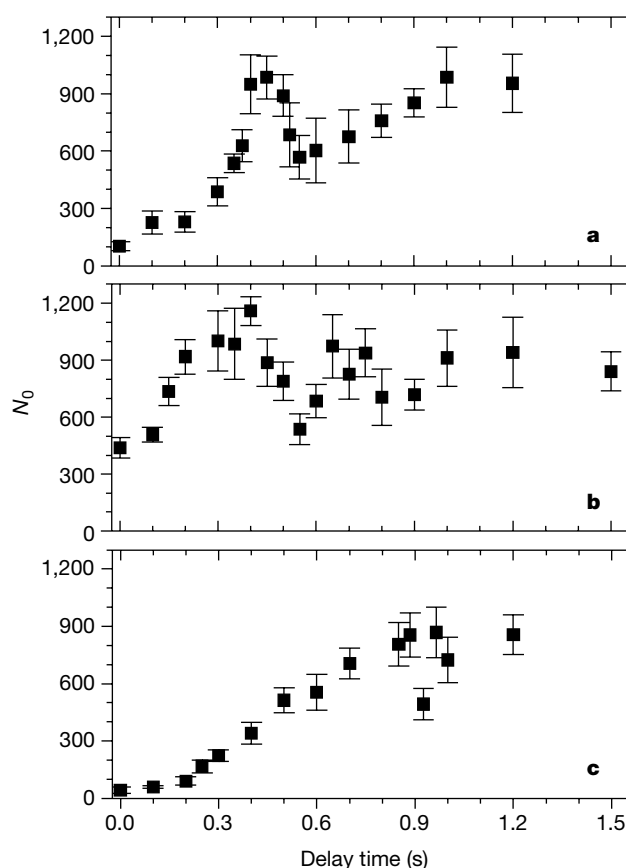


Figure 3 Condensate growth and collapse dynamics. Each data point is the mean of five measurements for the same delay time following the light pulse, and the error bars are the standard deviation of the mean. **a**, N_0 is reduced to ~ 100 atoms using the two-photon light pulse. The reduction in N_0 after the first peak at 450 ms is a direct manifestation of collapse. **b**, The condensate is only partially dumped in order to speed up the initial growth. Subsequent maxima and minima are observed as the gas continues to undergo growth and collapse cycles while evolving towards thermal equilibrium. **c**, The light pulse duration is increased, so that the condensate is dumped completely, to within the experimental resolution. We note the slow turn-on of growth and the subsequent saturation in the growth rate (see text). For **a** and **c**, there is a 3-s delay following the microwave quench pulse before the light pulse; whereas in **b**, the delay is 5 s. Additionally, for **a** and **c**, $N = (7 \pm 1) \times 10^4$ atoms and $T = (170 \pm 15)$ nK immediately before the light pulse, while for **b**, $N = (1.0 \pm 0.1) \times 10^5$ atoms and $T = (200 \pm 20)$ nK. For each individual image, the statistical uncertainty in N_0 is ± 60 atoms, while the systematic uncertainty, dominated by uncertainties in the imaging system, is $\pm 20\%$ ¹³.

Destructive phase-contrast imaging is used to determine N_0 , the total number of atoms N , and their temperature T (ref. 15).

Following a delay of several seconds after the microwave quench pulse, the condensate is dumped by a light pulse consisting of two co-propagating laser beams whose frequency difference is tuned to resonance between the collisional state of two free atoms and a vibrational level of the diatomic molecule Li_2 , as shown in Fig. 1. Once in the molecular state, the laser of frequency ω_2 can stimulate a single-photon transition to the intermediate level ν' , which may spontaneously decay, most probably into a state of two energetic atoms that will escape the trap. This method for removing atoms is very energy-specific because the observed two-photon linewidth of ≈ 500 Hz is much less than the ~ 5 kHz thermal energy spread of the trapped atoms. In particular, the condensate may be selectively removed without significantly affecting the remaining atoms. This is demonstrated in Fig. 2, where both the measured condensate fraction and the fraction of total atoms are plotted as a function of the duration of the light pulse. The two-photon spectroscopy of this work is similar to that performed previously in a non-quantum-degenerate gas of lithium atoms at $T \approx 1$ mK (ref. 17), and in a Bose condensate of Rb atoms¹⁸.

After the light pulse, the gas is allowed to evolve freely for a certain time, at which point a destructive measurement of N_0 is made. Figure 3a shows the dynamical evolution of the condensate after a

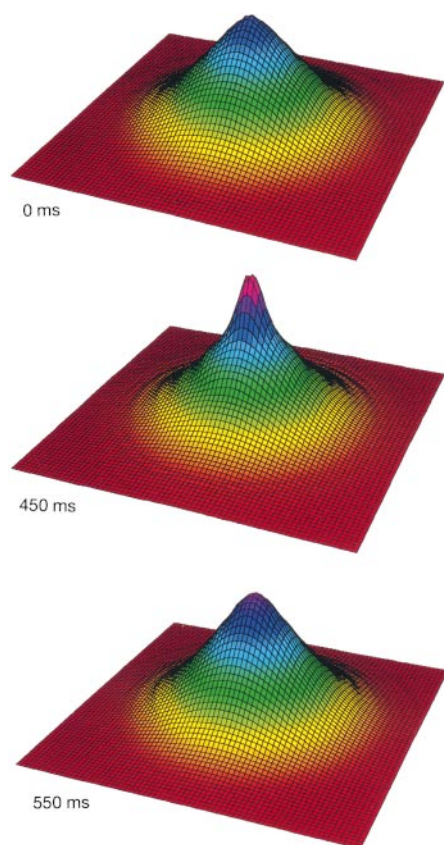


Figure 4 Phase-contrast images. These images were selected from those used to construct Fig. 3a. The upper image corresponds to data immediately after the dump pulse (0 ms), the middle image to the peak of condensate growth (450 ms), and the lower image to the first collapse (550 ms). The fitted values of N_0 are 40, 1,210 and 230 atoms for the upper, middle and lower images, respectively. The displayed images result from angle-averaging the actual images about the probe laser propagation axis and the signal height is proportional to the column density of the atom cloud integrated along this direction⁴. The image area corresponds to a square of sides $85 \mu\text{m}$. A movie of the growth and collapse was produced using representative time-ordered images from the data of Fig. 3a, and is available for viewing at <http://atomcool.rice.edu/collapse.html>.

light pulse whose duration is adjusted to reduce N_0 to an initial value of ~ 100 atoms. N_0 increases immediately after the light pulse as the condensate is fed by collisions between noncondensed thermal atoms, reaching a maximum value consistent with the expected upper limit of 1,250 atoms. A collapse is clearly indicated by the subsequent reduction in N_0 . After the collapse, N_0 grows again, as the gas is not yet in thermal equilibrium. Figure 4 shows representative phase-contrast images taken from the data in Fig. 3a for the specified delay times. The central peak, most clearly visible at 450 ms delay, corresponds to the condensate.

The condensate growth rate may be adjusted by varying the duration of the light pulse. By reducing the duration, fewer atoms are removed from both the condensate and from the low-energy thermal atoms that directly contribute to condensate growth, and consequently the growth rate increases. This is shown in Fig. 3b, where two secondary peaks are now discernible. For Fig. 3c, the light pulse duration is lengthened, causing more atoms to be removed, and slowing the rate of growth.

Each data point in Fig. 3 is the mean of five separate measurements of N_0 . Consequently, these data represent an average of many trajectories whose initial phase and rate of growth differ slightly. To analyse the results, we numerically simulate the collisional redistribution of atoms over the energy states of the trap using the quantum Boltzmann equation, as described in detail elsewhere⁹. The coloured curves shown in Fig. 5 are a sample of simulated trajectories which include the effect of the microwave quench pulse and the light pulse. The variation in condensate growth following a light pulse is mainly the result of slight differences in initial conditions that lead to variations in the energy distribution of atoms in the trap. Additionally, the stochastic nature of the collapse process, which causes each collapse to occur at a slightly different value of N_0 , contributes to dephasing of different trajectories. The heavy black line represents the average of 40 trajectories obtained by running the simulation several times with slightly different initial conditions. In the simulations, N_0 is set to 100 atoms immediately following the light pulse in order to provide direct comparison with the data in Fig. 3a. Additionally, N_0 is set to 200 atoms immediately after a collapse in order to achieve the best agreement with our previous statistical studies¹³. The only adjustable parameter in the simulations was the fraction of thermal atoms lost within the spectral width of the two-photon transition.

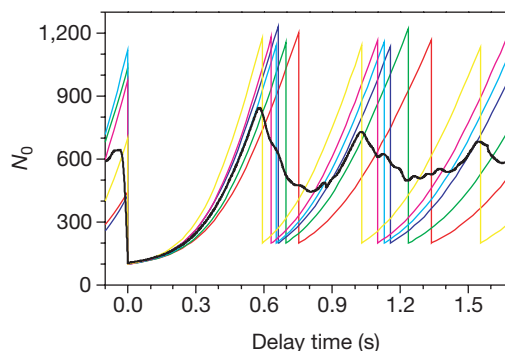


Figure 5 Simulation of condensate dynamics. The coloured curves show individual trajectories generated by a numerical quantum Boltzmann simulation, which models the kinetics of evaporative cooling and thermal equilibration. Individual trajectories were generated by running the simulation for 40 different initial values of N spaced evenly between 2.4×10^8 and 2.6×10^8 atoms. The resulting values of N just before the light pulse range between 6×10^4 and 8×10^4 atoms, in agreement with the range of N observed in the data. The heavy solid line is the mean of these 40 trajectories. The effect of the light pulse on the thermal atoms is simulated by instantaneously reducing the population of atoms within a 500 Hz wide Lorentzian energy band, consistent with the observed spectral width. For the simulation shown, 90% of the thermal atoms at the Lorentzian peak are lost.

The simulation results agree well with the data in several respects. The data in Fig. 3b show that each subsequent peak following the initial growth is slightly lower, as the trajectories corresponding to each individual measurement dephase from one another. This dephasing causes the maxima (minima) to occur at smaller (larger) values of N_0 than for any individual trajectory. This behaviour is seen in the simulation average shown in Fig. 5, confirming our understanding of the role of averaging in these measurements.

Condensate growth should be affected by the quantum statistical effect known as Bose enhancement: the condensate growth rate scales with the occupation number N_0 . This would lead to an exponentially increasing growth rate, which is, however, neither observed in the data nor predicted by the quantum Boltzmann equation model. For the conditions shown in Fig. 3c, for example, both the model and the data show a saturation in the growth rate when N_0 becomes larger than ~ 150 atoms. An analysis of the distribution of atoms among the trap energy levels in the model simulations indicates that the condensate is fed mostly by collisions between atoms with the lowest energies. Further, this analysis indicates that when the low-energy population is depleted, the growth is limited by the rate for non-Bose-enhanced 'trickle-down' collisions between high-energy atoms. Owing to our high sensitivity to small values of N_0 , the very early stages of condensate growth following the dump can be observed, illuminating this subtle, yet important, departure from the pure Bose enhancement prediction. The first observation of condensate growth¹⁹ could not detect small values of N_0 and therefore was not sensitive to the initial stages of growth.

To our knowledge, the data presented here are the first direct observations of the growth and collapse of Bose–Einstein condensates with attractive interactions. By enabling the preparation of condensates with a desired mean number of atoms with minimal perturbation to the noncondensed atoms, the two-photon technique provides a new way of studying the dynamics of this non-equilibrium and stochastic quantum system. □

Received 12 September; accepted 24 October 2000.

1. Stoof, H. T. C. Atomic Bose gas with a negative scattering length. *Phys. Rev. A* **49**, 3824–3830 (1994).
2. Bradley, C. C., Sackett, C. A., Tollett, J. J. & Hulet, R. G. Evidence of Bose–Einstein condensation in an atomic gas with attractive interactions. *Phys. Rev. Lett.* **75**, 1687–1690 (1995).
3. Ruprecht, P. A., Holland, M. J., Burnett, K. & Edwards, M. Time-dependent solution of the nonlinear Schrödinger equation for Bose-condensed trapped neutral atoms. *Phys. Rev. A* **51**, 4704–4711 (1995).
4. Bradley, C. C., Sackett, C. A. & Hulet, R. G. Bose–Einstein condensation of lithium: observation of limited condensate number. *Phys. Rev. Lett.* **78**, 985–989 (1997).
5. Kagan, Y., Shlyapnikov, G. V. & Walraven, J. T. M. Bose–Einstein condensation in trapped atomic gases. *Phys. Rev. Lett.* **76**, 2670–2673 (1996).
6. Shuryak, E. V. Bose condensate made of atoms with attractive interaction is metastable. *Phys. Rev. A* **54**, 3151–3154 (1996).
7. Stoof, H. T. C. Macroscopic quantum tunneling of a Bose condensate. *J. Stat. Phys.* **87**, 1353–1366 (1997).
8. Ueda, M. & Leggett, A. J. Macroscopic quantum tunneling of a Bose–Einstein condensate with attractive interactions. *Phys. Rev. Lett.* **80**, 1576–1579 (1998).
9. Sackett, C. A., Stoof, H. T. C. & Hulet, R. G. Growth and collapse of a Bose condensate with attractive interactions. *Phys. Rev. Lett.* **80**, 2031–2034 (1998).
10. Kagan, Y., Muryshv, A. E. & Shlyapnikov, G. V. Collapse and Bose–Einstein condensation in a trapped Bose gas with negative scattering length. *Phys. Rev. Lett.* **81**, 933–937 (1998).
11. Houbiers, M. & Stoof, H. T. C. Stability of Bose condensed atomic ^7Li . *Phys. Rev. A* **54**, 5055 (1996).
12. Chandrasekhar, S. *An Introduction to the Study of Stellar Structure* (Dover, New York, 1957).
13. Sackett, C. A., Gerton, J. M., Welling, M. & Hulet, R. G. Measurements of collective collapse in a Bose–Einstein condensate with attractive interactions. *Phys. Rev. Lett.* **82**, 876–879 (1999).
14. Andrews, M. R. *et al.* Direct, nondestructive observation of a Bose condensate. *Science* **273**, 84–87 (1996).
15. Sackett, C. A., Bradley, C. C., Welling, M. & Hulet, R. G. Bose–Einstein condensation of lithium. *Appl. Phys. B* **65**, 433–440 (1997).
16. Cornish, S. L., Claussen, N. R., Roberts, J. L., Cornell, E. A. & Wieman, C. E. Stable ^{85}Rb Bose–Einstein condensates with widely tunable interactions. *Phys. Rev. Lett.* **85**, 1795–1798 (2000).
17. Abraham, E. R. I., McAleer, W. I., Sackett, C. A. & Hulet, R. G. Spectroscopic determination of the s-wave scattering length of lithium. *Phys. Rev. Lett.* **74**, 1315–1318 (1995).
18. Wynar, R., Freeland, R. S., Han, D. J., Ryu, C. & Heinzen, D. J. Molecules in a Bose–Einstein condensate. *Science* **287**, 1016–1019 (2000).
19. Miesner, H.-J. *et al.* Bosonic stimulation in the formation of a Bose–Einstein condensate. *Science* **279**, 1005–1007 (1998).

Acknowledgements

We thank C.A. Sackett for help with the quantum Boltzmann simulation. This work was supported by the US National Science Foundation, the National Aeronautics and Space Administration, the Office of Naval Research and the Welch Foundation.

Correspondence and requests for materials should be addressed to R.G.H. (e-mail: randy@atomcool.rice.edu).

Direct observation of molecular cooperativity near the glass transition

E. Vidal Russell & N. E. Israeloff

Department of Physics and Center for Interdisciplinary Research on Complex Systems, Northeastern University, Boston, Massachusetts 02115, USA

The increasingly sluggish response of a supercooled liquid as it nears its glass transition¹ (for example, refrigerated honey) is prototypical of glassy dynamics found in proteins, neural networks and superconductors. The notion that molecules rearrange cooperatively has long been postulated² to explain diverging relaxation times and broadened (non-exponential) response functions near the glass transition. Recently, cooperativity was observed and analysed in colloid glasses³ and in simulations of binary liquids well above the glass transition⁴. But nanometre-scale studies of cooperativity at the molecular glass transition are lacking⁵. Important issues to be resolved include the precise form of the cooperativity and its length scale⁶, and whether the broadened response is intrinsic to individual cooperative regions, or arises only from heterogeneity^{7–9} in an ensemble of such regions. Here we describe direct observations of molecular cooperativity near the glass

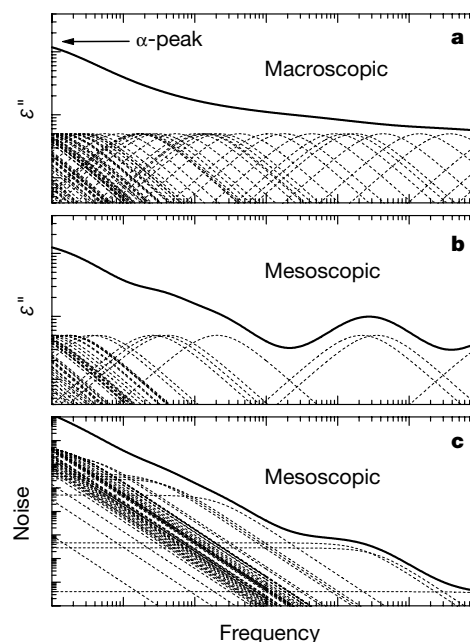


Figure 1 Heterogeneous scenario. **a**, The imaginary component of the dielectric susceptibility, ϵ'' , is assumed to arise from a superposition of Debye peaks. In a mesoscopic sample (**b**), the individual Debye peaks become apparent as spectral features, particularly in the high-frequency tail of the α -peak. In the thermal noise spectrum (**c**) the Debye peaks appear as Lorentzians, which also produce spectral features.

transition in polyvinylacetate (PVAc), using nanometre-scale probing of dielectric fluctuations. Molecular clusters switched spontaneously among two to four distinct configurations, producing random telegraph noise. Our analysis of these noise signals and their power spectra reveals that individual clusters exhibit transient dynamical heterogeneity and non-exponential kinetics.

Dielectric relaxation of a typical supercooled liquid slows with decreasing temperature and becomes non-exponential near the glass transition. The corresponding peak in the imaginary component of the dielectric susceptibility, $\epsilon''(\omega)$, becomes broader in frequency, ω , than a single relaxation-time Debye peak¹⁰. In the heterogeneous case, this broadened primary (α) response arises from cooperative regions² (clusters) of typical size, 2–4 nm (ref. 6), each with Debye-like response, but with characteristic relaxation times, τ , which differ from cluster to cluster as a result of disorder (see Fig. 1a). Whether heterogeneity is the sole source of broadening has been controversial. Dynamically selective experiments have found heterogeneity that is either short-lived^{7,8} or long-lived⁹ compared with the average relaxation time, whereas other experiments may support intrinsic broadening or homogeneity¹¹.

If it were practical to probe the susceptibility of an individual cooperative region within a glass, many of these questions might be addressed. Even a slightly larger volume containing a small number of such regions could be useful, because in the heterogeneous case, the responses of individual clusters will be centred on different frequencies, producing distinguishable features in the susceptibility spectrum. These deviations from macroscopic response would be most pronounced in the high-frequency tail of the primary (α)-susceptibility peak, where scaling behaviour is often observed¹⁰, and where very few clusters should be contributing (Fig. 1b). One potential problem in probing such a small volume is that fluctuations (noise) should become relatively large. Fortunately, the noise spectrum contains the same information as the susceptibility spectrum and can be derived from it using the fluctuation–dissipation theory, as long as the sample is in equilibrium¹². Debye-shaped susceptibility features produce Lorentzian-shaped features in the noise spectrum (Fig. 1c). If observed, these spectral

features would provide direct evidence for heterogeneity and cooperativity and make new details of glassy dynamics accessible.

By sensing local electrostatic forces, non-contact atomic force microscopy (AFM) techniques can be used to measure local dielectric properties¹³. Recently, dielectric relaxation¹⁴ and spontaneous thermal dielectric noise¹⁵ were studied in glassy polymers with non-contact AFM techniques. In the present experiment, we employed a noise measurement scheme¹⁶ (see Fig. 2a), which used a small cantilever with a sharp (50 nm radius) conductive tip, mounted in a temperature-controlled can within a vacuum chamber. The cantilever is driven at its resonance frequency (~ 160 kHz), via a piezoelectric and a self-driven oscillator circuit. Samples were 0.5- μm thick PVAc films (glass transition: $T_g = 308$ K), which were spin-coated onto a metal substrate and annealed^{14,15}. When a voltage bias is applied between tip and substrate, fluctuations in the sample polarization beneath the tip produce proportional variations in the cantilever resonance frequency¹⁶ which are detected using frequency modulation techniques¹⁷. For small tip heights, z , above the sample surface, the resonance frequency is most sensitive to a small region directly beneath the tip. For $z = 30$ nm and a bias of 4 V, we numerically calculated the maximum electric field ($E_{\text{max}} = 8.5 \times 10^6 \text{ V m}^{-1}$), the effective probed volume ($\Omega \approx 2 \times 10^{-17} \text{ cm}^3$) and depth (40 nm). This depth exceeds the length scales below which surface effects dominate the dynamics¹⁸.

With tip height fixed using feedback to reset the resonance frequency between each set of data points (every 16–65 s), time series of PVAc polarization fluctuations were recorded at various temperatures slightly below T_g and Fourier analysed to obtain a power spectrum. Spectra averaged over long periods (≈ 1 day) showed the expected power law¹⁵, $S(f) \approx f^{-\gamma}$. However, anomalous transient deviations were observed. An example is shown in Fig. 2b. Successive spectra, each measured over a one-hour period, show significant changes in shape. Most striking is the growth and subsequent reduction of a spectral feature, which is similar in shape to a slightly broadened Lorentzian (also shown). Similar spectral features, which persisted for up to a few hours, were seen many times in the temperature range 298–302 K, and in the frequency range 0.05–25 Hz in several samples. These spectral features provide convincing evidence for heterogeneity, which is, however, transient. Analysis of the average lifetime of these heterogeneities was carried out by studying the time autocorrelation function of the local spectral exponent ($-\delta \ln(S)/\delta \ln(f)$)¹⁵. As shown in Fig. 3, the heterogeneity lifetime was comparable at each temperature to the

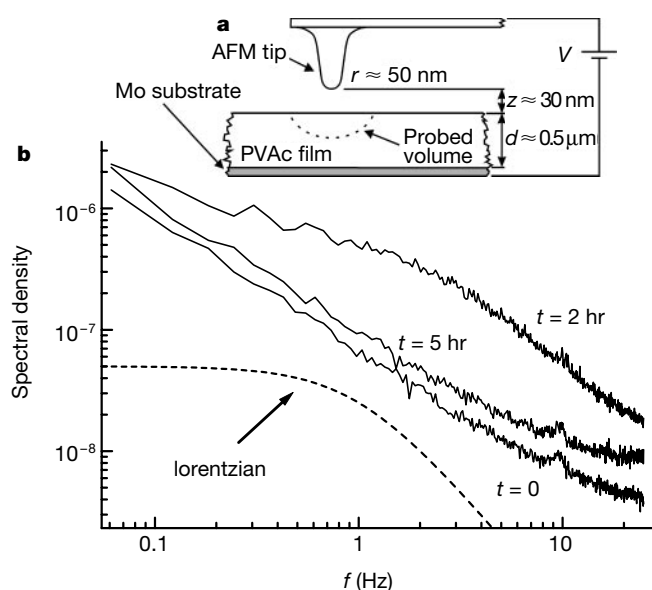


Figure 2 Noise spectra for PVAc at $T = 299$ K. **a**, The atomic force microscope (AFM) measurement set-up showing the cantilever and tip with radius, r , held at a height, z , above the surface of a PVAc film with thickness, d . **b**, Spectral density versus frequency. Successive polarization noise spectra show the appearance and subsequent disappearance of a spectral feature. Each spectrum is a one-hour average with the relative starting time indicated. The dashed line shows a Lorentzian spectrum.

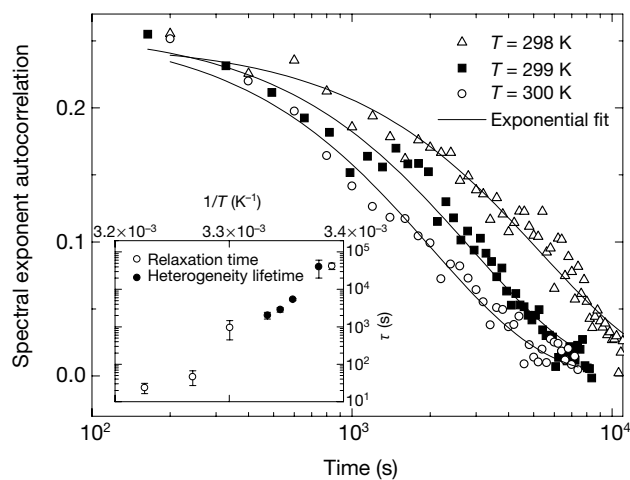


Figure 3 Lifetime of heterogeneities. The autocorrelation function of the local power spectral exponent¹⁵ is shown for three temperatures with exponential fits and characteristic times. Inset, heterogeneity lifetime and α -relaxation time¹⁴ are plotted versus inverse temperature.

measured¹⁴ average (α) relaxation time. This result demonstrates a link between these (~ 3 – 4 decades) faster processes and the primary relaxation, and matches the behaviour of heterogeneities found at higher temperatures and closer to the α -peak frequency⁸. We note that the β peak is 100 times smaller and 10 decades higher in frequency¹⁹ and is therefore unlikely to be relevant to these measurements.

Associated with the appearance of pronounced spectral features, we frequently observed random telegraph switching (RTS) among two to four discrete levels in the polarization time series. See Fig. 4. Like the spectral features, the RTS features were in all cases transient,

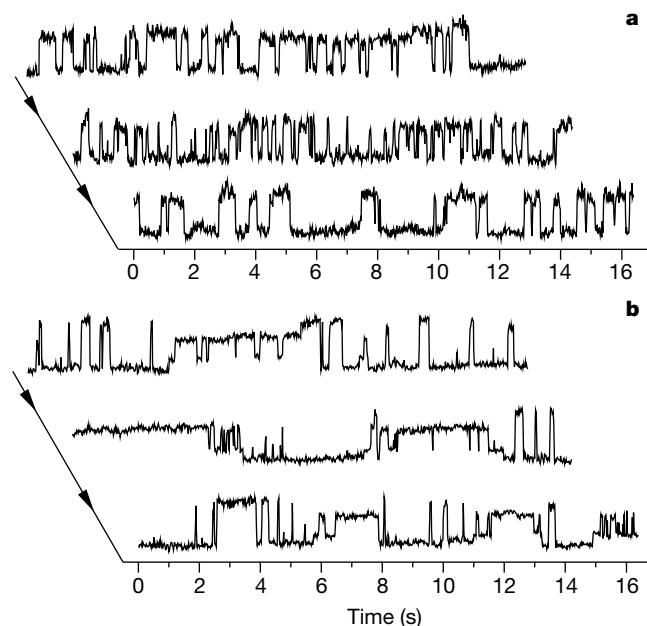


Figure 4 Time series of polarization. Two 50-s series showing random telegraph switching (RTS) noise. **a**, At 299 K, the rate of switching between two levels appears modulated. **b**, An example of a four-state switching at 300 K is shown.

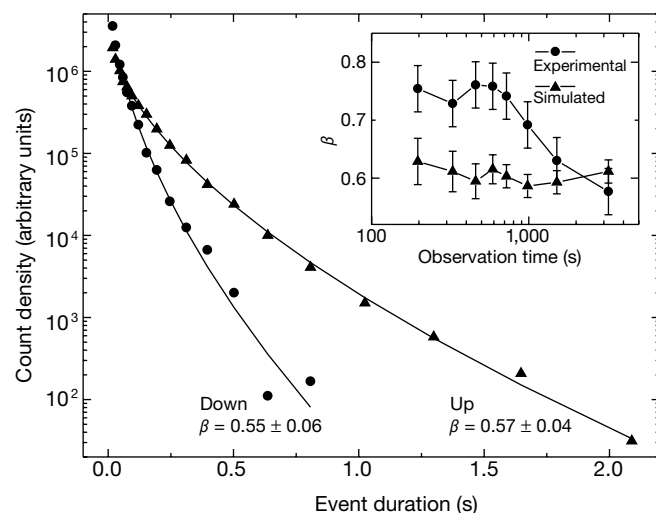


Figure 5 RTS distributions. Histograms of event durations for upper and lower states of a long-lived two-state RTS observed for 3,200 s at 299 K are shown (points) with stretched exponential fit (solid line). An exponential would appear as a straight line on this plot. The fitting parameters are indicated. Inset, average stretching exponent versus observation time for the upstate data (circles) and a simulated intrinsic stretched exponential RTS (triangles).

persisting for times ranging from seconds (a few cycles) to at most a few hours (thousands of cycles), and in some cases appearing and disappearing several times. The average switching rates of different observed RTS varied over the observable range (factor of ~ 100), clearly showing dynamical heterogeneity. Often, as seen in Fig. 4a, RTS activity or rates were modulated in time, sometimes by discrete transitions, possibly owing to other RTS.

Recent nuclear magnetic resonance studies²⁰ suggest that orientational relaxation in PVAc occurs via a random walk of $\sim 10^\circ$ angular jumps, with a typical jump time shorter than the α relaxation time. The RTS such as those in Fig. 4b, with at least four repeatedly visited levels, can be modelled by a cluster whose orientation makes moderately sized jumps in a multiple-well orientational potential. In support of this model, we note that changes in the applied electric field produced small but reproducible changes in the RTS level occupancy probabilities, consistent with changes in cluster dipole-moment or dipole-moment orientation (E.V.R. and N.E.L., unpublished work). These were used to infer cluster dipole moments which were found to be several times the dipole moment of a single PVAc monomer and were compatible with recent PVAc experiments⁶ and theory²¹ that found cluster sizes of 30–90 monomers.

In order to analyse the RTS kinetics, distributions of time durations were extracted from individual long-lived RTS series with two dominant levels. For a thermally activated two-state process, the time durations, t , spent in each configuration would be expected to be exponentially distributed, for example: $N(t) = N_0 \exp(-t/\tau)$. However, for all RTS studied (see Fig. 5), the distributions could be well fitted with a stretched exponential: $N(t) = N_0 \exp(-t/\tau)^\beta$, with $0.4 < \beta < 0.6$, similar to the values found for bulk dielectric relaxation¹⁴. The τ values varied widely (factor of 100) for different RTS, again demonstrating heterogeneity. One possibility is that this stretching arises from discrete modulation of the rates (see Fig. 4a) of an underlying exponential process. To test this idea, long RTS time series were sliced into shorter sections (for example, of 200 s) corresponding to observation times which are short compared to the α -relaxation time. For some RTS, distributions appeared to be exponential in selected shorter sections. As shown in the Fig. 5 inset, with increasing observation time the average value of β decreased for these RTS, indicating an evolution with time toward a more stretched response. Other RTS remain highly stretched even for observation times of 200 s, consistent with a more rapid modulation of rates. A simulated, intrinsically stretched exponential RTS showed time-independent β values (as shown) and no exponential short-time behaviour, indicating that these effects were statistically significant.

Thus, using a nanoscale probe of dielectric fluctuations in PVAc, we have observed spontaneous switching between discrete polarization levels near the glass transition. These results show directly that the idea of cooperativity is applicable to molecular glasses. Cooperativity takes the form of transient molecular clusters, a few nanometres in size, which reorient (up to thousands of times) between 2 to 4 (or more) distinct configurations, and may be modelled by thermal activation in a multiple-well orientational potential. Clusters had a wide range of characteristic switching rates, direct evidence that the broadening of the α -relaxation process arises from a heterogeneity. The lifetime of heterogeneities was comparable to the α -relaxation time. Individual cluster configurations exhibited stretched exponential kinetics, with some showing a tendency toward exponential kinetics for observation times that were short compared with the α -relaxation time. This behaviour was linked with the observed discrete modulations of switching rates, which might be understood in terms of a slowly fluctuating environment²². □

Received 10 July; accepted 27 October 2000.

1. Angell, C. A. Formation of glasses from liquids and biopolymers. *Science* **267**, 1924–1939 (1995).

2. Adam, G. & Gibbs, J. H. On the temperature dependence of cooperative relaxation properties in glass forming liquids. *J. Chem. Phys.* **43**, 139–146 (1965).
3. Weeks, E. R., Crocker, J. L., Levitt, A. C., Schofield, A. & Weitz, D. A. Three-dimensional direct imaging of structural relaxation near the colloidal glass transition. *Science* **287**, 627–629 (2000).
4. Benmounem, C., Donati, C., Baschnagel, J. & Glotzer, S. C. Growing range of correlated motion in a polymer melt on cooling towards the glass transition. *Nature* **399**, 246–249 (1999).
5. Binder, K., Baschnagel, J., Kob, W. & Paul, W. Glass physics: still not transparent. *Phys. World* **12**, 54 (1999).
6. Tracht, U. *et al.* Length scale of dynamic heterogeneities at the glass transition determined by multidimensional nuclear magnetic resonance. *Phys. Rev. Lett.* **81**, 2728–2731 (1998).
7. Schiener, B., Bohmer, R., Loidl, A. & Chamberlin, R. V. Nonresonant spectral hole burning in the slow dielectric response of supercooled liquids. *Science* **274**, 752–754 (1996).
8. Bohmer, R. *et al.* Nature of the nonexponential primary relaxation in structural glasses probed by dynamically selective experiments. *J. Non-Cryst. Solids* **235**, 1–8 (1998).
9. Ediger, M. D. Spatially heterogeneous dynamics in supercooled liquids. *Annu. Rev. Phys. Chem.* **51**, (in the press).
10. Dixon, P. K., Wu, L., Nagel, S. R., Williams, B. D. & Carini, J. P. Scaling in the relaxation of a supercooled liquid. *Phys. Rev. Lett.* **65**, 1108–1111 (1990).
11. Arbe, A., Colmenero, J., Monkenbusch, M. & Richter, D. Dynamics of glass forming polymers: “homogeneous” vs. “heterogeneous” scenario. *Phys. Rev. Lett.* **81**, 590–593 (1998).
12. Grigera, T. S. & Israeloff, N. E. Observation of fluctuation-dissipation-theorem violations in a structural glass. *Phys. Rev. Lett.* **83**, 5038–5041 (1999).
13. Martin, Y., Abraham, D. W. & Wickramasinghe, H. K. High-resolution capacitance measurement and potentiometry by force microscopy. *Appl. Phys. Lett.* **52**, 1103–1105 (1988).
14. Walther, L. E., Israeloff, N. E., Vidal Russell, E. & Alvarez Gomariz, H. Mesoscopic scale dielectric relaxation at the glass transition. *Phys. Rev. B* **57**, R15112–R15115 (1998).
15. Vidal Russell, E., Walther, L. E., Israeloff, N. E. & Alvarez Gomariz, H. Nanometer scale dielectric fluctuations at the glass transition. *Phys. Rev. Lett.* **81**, 1461–1464 (1998).
16. Walther, L. E., Vidal Russell, E., Israeloff, N. E. & Alvarez Gomariz, H. Atomic force measurement of low frequency dielectric noise. *Appl. Phys. Lett.* **72**, 3223–3226 (1998).
17. Albrecht, T. R., Grütter, P., Horne, D. & Rugar, D. Frequency modulation detection using high-Q cantilevers for enhanced force microscope sensitivity. *J. Appl. Phys.* **69**, 668–673 (1991).
18. Forrest, J. A., Dalnoki-Veress, K., Stevens, J. R. & Dutcher, J. R. Effect of free surfaces on the glass transition temperature of thin polymer films. *Phys. Rev. Lett.* **77**, 2002–2005 (1996).
19. McCrum, N. G., Read, B. E. & Williams, G. *Elastic and Dielectric Effects In Polymeric Solids* 302–305 (Dover, New York, 1991).
20. Tracht, U., Heuer, A., Spiess, H. W. Geometry of reorientational dynamics in supercooled poly(vinyl acetate) studied by 2D NMR echo experiments. *J. Phys. Chem.* **111**, 3720–3727 (1999).
21. Xia, X. & Wolynes, P. G. Fragilities of liquids predicted from the random first order transition theory of glasses. *Proc. Natl Acad. Sci. USA* **97**, 2990–2994 (2000).
22. Wang, J. & Wolynes, P. Intermittency of single molecule reaction dynamics in fluctuating environments. *Phys. Rev. Lett.* **74**, 4317–4320 (1995).

Acknowledgements

This research was supported by the NSF Division of Materials Research, and the Petroleum Research Fund administered by the American Chemical Society. We thank M. D. Ediger for helpful discussions and K. Sinnathamby for assistance.

Correspondence and requests for materials should be addressed to N.E.I. (e-mail: Israeloff@neu.edu).

Evidence for decoupling of atmospheric CO₂ and global climate during the Phanerozoic eon

Ján Veizer*, Yves Godderis† & Louis M. François†

* Institut für Geologie, Mineralogie und Geophysik, Ruhr Universität, 44780 Bochum, Germany, and Ottawa-Carleton Geoscience Centre, University of Ottawa, Ottawa, ON K1N 6N5, Canada

† Laboratoire de Physique Atmosphérique et Planétaire, Université de Liège, 4000 Liège, Belgium

Atmospheric carbon dioxide concentrations are believed to drive climate changes from glacial to interglacial modes¹, although geological^{1–3} and astronomical^{4–6} mechanisms have been invoked as ultimate causes. Additionally, it is unclear^{7,8} whether the changes between cold and warm modes should be regarded as a global phenomenon, affecting tropical and high-latitude temperatures alike^{9–13}, or if they are better described as an expansion and

contraction of the latitudinal climate zones, keeping equatorial temperatures approximately constant^{14–16}. Here we present a reconstruction of tropical sea surface temperatures throughout the Phanerozoic eon (the past ~550 Myr) from our database¹⁷ of oxygen isotopes in calcite and aragonite shells. The data indicate large oscillations of tropical sea surface temperatures in phase with the cold–warm cycles, thus favouring the idea of climate variability as a global phenomenon. But our data conflict with a temperature reconstruction using an energy balance model that is forced by reconstructed atmospheric carbon dioxide concentrations¹⁸. The results can be reconciled if atmospheric carbon dioxide concentrations were not the principal driver of climate variability on geological timescales for at least one-third of the Phanerozoic eon, or if the reconstructed carbon dioxide concentrations are not reliable.

Calculated partial pressures of CO₂ in the atmosphere^{18–20} (*p*CO₂) are indeed relatively low for the Permo/Carboniferous and Cenozoic icehouse episodes¹, times of predominantly cold climates lasting tens of millions of years², but for the other two Phanerozoic icehouses—the late Ordovician/earliest Silurian and the late Jurassic/early Cretaceous—all global biogeochemical models^{19,20} predict high *p*CO₂ levels. In an attempt to resolve this discrepancy, particularly for the late Ordovician glaciation(s), theoretical models advocating the development of permanent high-latitude ice sheets²¹ at more than 10 times present-day CO₂ levels have been proposed. As an alternative, we present here experimental evidence that suggests large variations in tropical sea surface temperatures (SSTs; up to 9 °C) between the peaks of greenhouse/icehouse modes.

The testing of model predictions on Phanerozoic timescales has been hampered by the lack of high-resolution isotope databases for ancient sea water, as reflected in (bio)chemical sediments. This has been particularly the case for oxygen isotopes, where the depletion in ¹⁸O for progressively older sediments was mostly considered to be a product of post-depositional diagenetic overprint. However, new databases have been published¹⁷ recently—based on ~4,500 low-Mg calcitic and 125 aragonitic shells—for δ¹⁸O, δ¹³C and ⁸⁷Sr/⁸⁶Sr, establishing the baseline for Phanerozoic δ¹⁸O (Fig. 13 in ref. 17). Most studied samples originate from the photic zone of palaeotropical seas (30 °S–30 °N). The entire datasets are available at http://www.science.uottawa.ca/geology/isotope_data. Reference 17, because of the huge number of measurements, dealt only with the documentation of data and their quality, and with the arguments for the primary nature of the δ¹⁸O record.

The secular trend for the δ¹⁸O values of Phanerozoic carbonates, Fig. 13 in ref. 17, consists of a long-term rising linear trend that exceeds the duration of planetary greenhouse/icehouse modes. This trend is tectonically controlled, a proposition supported by the strong correlation of δ¹⁸O with the ⁸⁷Sr/⁸⁶Sr signal for the Phanerozoic sea water (Figs 16 and 17 in ref. 17). Sr isotopes are a proxy for terrestrial tectonic processes, because seawater ⁸⁷Sr/⁸⁶Sr is controlled principally by inputs of Sr from rivers (weathering of ⁸⁷Sr-enriched continental crust) and from hydrothermal circulation cells at mid-ocean ridges (interaction with ⁸⁷Sr-depleted basalts). Such long-term tectonic signals have to be subtracted from the superimposed higher-order oscillations before any discussion of climatic consequences, and we therefore detrend the data by removing the least-squares linear fit.

The first-order δ¹⁸O oscillations around the least-squares fit correlate well with the palaeoclimate as established from sedimentological criteria¹. The icehouses usually coincide with ¹⁸O-enriched values and greenhouses with ¹⁸O-depleted values (Fig. 1). We propose therefore that the Phanerozoic δ¹⁸O oscillations reflect variations in SSTs. Future improvements on the Phanerozoic database may result in amelioration of the amplitudes, or in partial temporal shifts of some of the peaks. This is particularly the case for the Neogene interval, where the running mean is forced to higher values by the inclusion of some isotopically heavy planktonic

foraminifera from the Northern Pacific Ocean. Subsequently, the running mean declines sharply to modern values for tropical calcite. Nonetheless, considering the size and quality of this database, it is unlikely that any future refinements will unravel the overall correlation between palaeoclimate and $\delta^{18}\text{O}$.

The proposition that the detrended $\delta^{18}\text{O}$ oscillations reflect climatically controlled SSTs is also supported by a correlation with the palaeolatitudinal distribution of ice-rafted debris (PIRD) as a function of geological age²² (Fig. 1). In this study, all data were re-calibrated to the timescale of ref. 23. At first, a simple statistical analysis that was based on the detrended $\delta^{18}\text{O}$ (running mean at 10/50 Myr; see Fig. 1 legend for nomenclature) and on the PIRD signal interpolated every 10 Myr at extant PIRD (29 points, with intervals without ice-rafted debris removed from the analysis), yielded only a

mediocre correlation coefficient (0.55). This is not surprising, considering the coarse time resolution of the PIRD database compared to the $\delta^{18}\text{O}$ signal. We then performed a decomposition in Fourier series of the detrended $\delta^{18}\text{O}$ and PIRD signals. Both series display a dominant component at ~ 135 Myr, regardless of the resolution adopted for the $\delta^{18}\text{O}$ signal (Fig. 2). Thus the result is not dependent on the parameters, such as window size and time-step, that were adopted for the running mean procedure. The maximal calculated dephasing between the two signals is 9 Myr, which is not significant considering the coarse resolution of the PIRD. The coherence, a measure of correlation between two periodic signals, is 0.92. This correlation between past global climate and the detrended $\delta^{18}\text{O}$ signal is further corroborated by the frequency histograms of other glacial deposits (OGD), at present

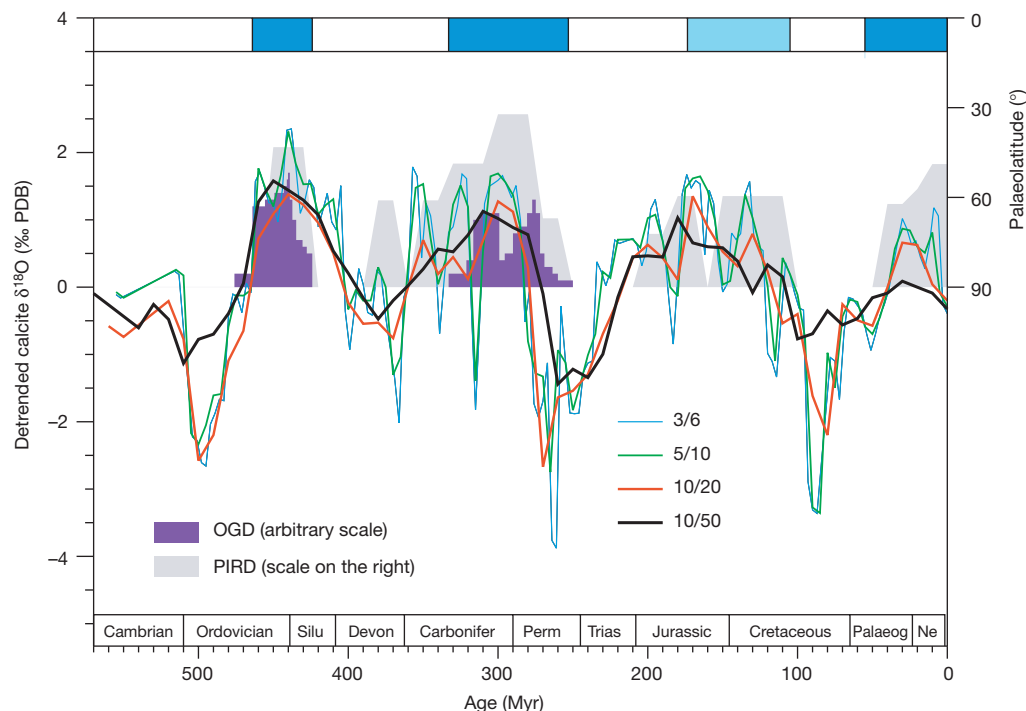


Figure 1 Detrended running means of $\delta^{18}\text{O}$ values of calcitic shells for the Phanerozoic. Detrending was achieved by subtracting the least-squares linear fit from the original data. 3/6, 5/10, 10/20 and 10/50 indicate running means at various temporal resolutions: for example, 3/6 means step 3 Myr, window 6 Myr. The palaeolatitudinal distribution of ice-rafted debris (PIRD)²² is on the right-hand vertical axis. The frequency histograms of other

glacial deposits (OGD)—such as tillites and glacial marine strata—for pre-Triassic times¹ are dimensionless. The blue bars at the top represent cool climate modes (icehouses) and the white bars the warm modes (greenhouses), as established from sedimentological criteria¹. The lighter blue shading for the Jurassic/Cretaceous icehouse reflects the fact that true polar ice caps have not been documented for this time interval.

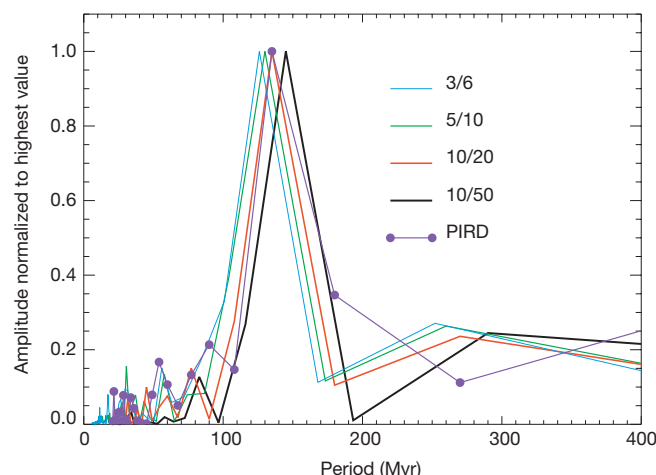


Figure 2 Periodogram of the Fourier series decomposition of the detrended $\delta^{18}\text{O}$ signal and of palaeolatitudes of ice-rafted debris (PIRD). See Fig. 1 for explanations of 3/6 to 10/

50. The PIRD result is based on 10 Myr interpolation. The curves are normalized to 1 for the highest value.

available only for the Ordovician/Silurian and Permo/Carboniferous glaciations¹ (Fig. 1).

From the above statistical study, we conclude that the observed oscillations of the detrended $\delta^{18}\text{O}$ signal are of climatic origin and reflect seawater temperatures. Because of the detrending, we shall discuss below the relative temperature deviations from the norm, and not the absolute temperatures.

Depending on temporal resolution, the detrended $\delta^{18}\text{O}$ oscillations have a magnitude of ~ 3 to $\sim 5\text{‰}$ between peaks of greenhouse and icehouse climates (Fig. 1). The coarser the resolution, the less the magnitude of the peaks. The greenhouse/icehouse oscillations of $\sim 3\text{‰}$ would indicate temperature variations of $\sim 13^\circ\text{C}$. However, some of this 3‰ range may be due to an ice-volume effect. Assuming no ice caps at the peaks of greenhouse climates, and assuming ice sheets of twice the present-day size for the peaks of icehouse climates, the ice-volume effect may account for about 2‰ of this range^{9,24}, thus reducing the actual temperature oscillations to only $\sim 4.5^\circ\text{C}$. Yet, the 3‰ is a reduced amplitude at 10/50 Myr resolution. By analogy with the Quaternary period, the ice effect influences seawater $\delta^{18}\text{O}$ on much shorter timescales and an amplitude of $\sim 4\text{‰}$ for the $\delta^{18}\text{O}$ (carbonate) is thus a more realistic estimate, a proposition clearly supported by the 3/6 to 10/20 curves (Fig. 1). If so, the residual after allowing for the ice-volume effect would still be $\sim 2\text{‰}$, or some 9°C .

The above interpretation of experimental data sheds some light on two issues that are at present subject to debate. (1) Most samples in the databases¹⁷ originate from shallow-water habitats of low-latitude, mostly palaeotropical seas. We therefore suggest that we are dealing with marked temperature variations that were of global nature and not solely events of higher latitudes. This assertion is in accord with the proposed stadial/interstadial tropical SST variations of $4\text{--}5^\circ\text{C}$ indicated by Quaternary proxy data^{9–13}, as well as with the stadial/interstadial $\delta^{18}\text{O}$ pattern within the Ordovician/Silurian icehouse (Fig. 15 in ref. 17). (2) The $\delta^{18}\text{O}$ -based tropical palaeotemperatures disagree with palaeotemperatures predicted from atmospheric $p\text{CO}_2$, the latter reconstructed from carbon isotope data for palaeosols and from the stomatal index of fossil leaves (compilations in ref. 20). All these reconstructions show high $p\text{CO}_2$ before 350 Myr ago, a marked decrease during the Permo/Carboniferous glaciations, high values in the Mesozoic, and a final decline

through Cenozoic times. We have used these reconstructions of atmospheric CO_2 as an input for an energy-balance climate model¹⁸. This model incorporates changes in palaeogeography, as well as the slow increase of about 5% in the solar constant during the Phanerozoic²⁵. For the present-day configuration, and a doubling of atmospheric CO_2 , the model produces an increase in the global mean temperature of 2.8°C ; this is at the lower end of the range of climate sensitivity to CO_2 . The IPCC 1995 report²⁶ gives a sensitivity of $2.8\text{--}5.2^\circ\text{C}$ for the atmosphere-only models, but a higher sensitivity would only compound the difficulties mentioned below. The simulations based on such a climate model yield temperatures that are in serious disagreement with the $\delta^{18}\text{O}$ -scaled tropical temperatures, particularly during Ordovician/Silurian and Jurassic/Cretaceous times. At these, mostly icehouse, times, the model simulations show tropical SSTs that were about $5 \pm 1^\circ\text{C}$ warmer than the present-day value, while the $\delta^{18}\text{O}$ -scaled SSTs were colder by about 2.5°C (Fig. 3) than the Phanerozoic norm. It is the latter that are in much better agreement with geological observations.

We note that the progressive ^{18}O depletion with age in the course of the Phanerozoic (Fig. 13 in ref. 17) has been claimed to be a reflection of: (1) increasing post-depositional alteration²⁷; (2) stratified oceans due to formation of warm, saline, deep waters²⁸; and (3) advancing cooling of the oceans²⁹. These arguments have been discussed in detail, and discounted, in ref. 17. Nevertheless, and leaving aside these counter arguments, we consider that (1)–(3) above are not the cause of the effects that we report here. For a post-depositional trend (1), the magnitude of the greenhouse/icehouse $\delta^{18}\text{O}$ oscillations would have to be inherited from the original anomalies. As these would be diminished in the course of alteration, the original anomalies would have to have been even larger than the observed ones, thus only compounding the problem. The alternative (2) would preferentially be associated with warm episodes, as opposed to glacial times, and for the Ordovician could account for no more than $\sim 0.9\text{‰}$ ^{18}O depletion in surface waters²⁸. Finally, for alternative (3), one-half of the entire trend would have to be generated by progressive cooling if the Ordovician $\delta^{18}\text{O}$ -scaled temperatures were to match the results of the energy-balance model. Yet, the same assumption would also generate Cambrian and Devonian tropical SSTs respectively 27 and 12°C warmer than the Phanerozoic mean, and analogous problems would arise

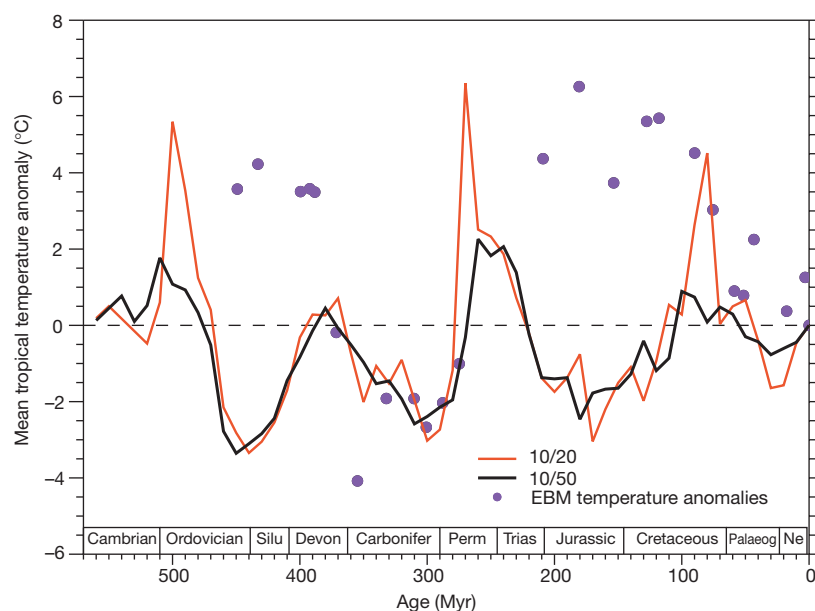


Figure 3 Tropical surface palaeotemperature anomalies calculated by the energy-balance climate model, and sea surface temperatures (SSTs) inferred from the $\delta^{18}\text{O}$ data. Filled circles, model predictions; red and black lines, SSTs calculated assuming the existence of ice sheets of twice the present-day size at the peaks of icehouse climates,

with sea water $\delta^{18}\text{O}$ scaled to -0.7‰ SMOW at times with no ice caps. The palaeothermometer of ref. 30 is applied, assuming that the detrended baseline seawater $\delta^{18}\text{O}$ value is 0‰ SMOW. Further explanations as in Fig. 1.

before, and following, the Jurassic/Cretaceous icehouse. Considering the similarity of faunal assemblages, in our case brachiopods, this is an unpalatable proposition¹⁷.

Three possible implications of our findings are: (1) the reconstructed past CO₂ levels are (partially) incorrect; (2) the role of pCO₂ as the main driving force of past global (long-term) climate changes is questionable, at least during two of the four main climate modes of the Phanerozoic; and (3) climate models, which include numerous parametrizations, are calibrated to the present (an interstadial in an icehouse climate), and may thus be unable to reproduce correctly the past climate modes. We hope that our results will stimulate further work on these three issues. □

Received 7 July; accepted 2 October 2000.

1. Frakes, L. A., Francis, J. E. & Syktus, J. I. *Climate Modes of the Phanerozoic* (Cambridge Univ. Press, Cambridge, 1992).
2. Fischer, A. G. in *Climate in Earth History* (eds Berger, W. H. & Crowley, J. C.) 97–104 (Nat'l Acad. Sci., Washington DC, 1982).
3. Worsley, T. R., Nance, R. D. & Moody, J. B. Tectonic cycles and the history of the Earth's biogeochemical and paleoceanographic record. *Paleoceanography* **1**, 233–263 (1986).
4. Steiner, J. & Grillmair, E. Possible galactic causes for periodic and episodic glaciations. *Geol. Soc. Am. Bull.* **84**, 1003–1018 (1973).
5. Williams, G. E. Possible relation between periodic glaciation and the flexure of the galaxy. *Earth Planet. Sci. Lett.* **26**, 361–369 (1975).
6. Hays, J. D., Imbrie, J. & Shackleton, N. J. Variations in the Earth's orbit: pacemaker of the ice ages. *Science* **194**, 1121–1132 (1976).
7. Crowley, T. J. Pleistocene temperature changes. *Nature* **371**, 664 (1994).
8. Broecker, W. S. Glacial climate in the tropics. *Science* **272**, 1902–1904 (1996).
9. Guilderson, T. P., Fairbanks, R. G. & Rubenstone, J. L. Tropical temperature variations since 20,000 years ago: modulating interhemispheric climate change. *Science* **263**, 663–665 (1994).
10. Stute, M. et al. Cooling of tropical Brazil (5°C) during the last glacial maximum. *Science* **269**, 379–383 (1995).
11. Rind, D. & Peteet, D. Terrestrial conditions at the last glacial maximum and CLIMAP sea-surface temperature estimates: are they consistent? *Quat. Res.* **24**, 1–22 (1985).
12. Thompson, L. G. et al. Late glacial stage and Holocene tropical ice core records from Huascarán, Peru. *Science* **269**, 46–50 (1995).
13. Beck, J. W., Récy, J., Taylor, F., Edwards, R. L. & Gabioc, G. Abrupt changes in early Holocene tropical sea surface temperature derived from coral records. *Nature* **385**, 705–707 (1997).
14. CLIMAP Project Members *Map and Chart Series MC-36* (Geological Society of America, Boulder, Colorado, 1981).
15. Bush, A. B. G. & Philander, S. G. H. The role of ocean-atmosphere interactions in tropical cooling during the Last Glacial Maximum. *Science* **279**, 1341–1344 (1998).
16. Crowley, T. J. Climate SSTs re-visited. *Clim. Dyn.* **16**, 241–255 (2000).
17. Veizer, J. et al. ⁸⁷Sr/⁸⁶Sr, ^δ13C and ^δ18O evolution of Phanerozoic seawater. *Chem. Geol.* **161**, 59–88 (1999).
18. François, L. M. & Walker, J. C. G. Modelling the Phanerozoic carbon cycle and climate: constraints from the ⁸⁷Sr/⁸⁶Sr isotopic ratio of seawater. *Am. J. Sci.* **292**, 81–135 (1992).
19. Berner, R. A. GEOCARB II: a revised model for atmospheric CO₂ over Phanerozoic time. *Am. J. Sci.* **294**, 56–91 (1994).
20. Berner, R. A. The carbon cycle and CO₂ over Phanerozoic time: the role of land plants. *Phil. Trans. R. Soc. Lond. B.* **353**, 75–82 (1998).
21. Gibbs, M. T., Bice, K. L., Barron, E. J. & Kump, L. R. in *Warm Climates in Earth History* (eds Huber, B. T., MacLeod, K. G. & Scott, L. W.) 386–422 (Cambridge Univ. Press, Cambridge, 2000).
22. Frakes, L. A. & Francis, J. E. A guide to Phanerozoic cold polar climates from high-latitude ice-raffing in the Cretaceous. *Nature* **333**, 547–549 (1988).
23. Harland, W. B. et al. *A Geologic Timescale* (Cambridge Univ. Press, Cambridge, 1990).
24. Savin, S. M. The history of the earth's surface temperature during the past 100 million years. *Annu. Rev. Earth Planet. Sci.* **5**, 319–355 (1977).
25. Endal, A. S. & Sofia, S. Rotation in solar-type stars, I, Evolutionary models for the spin-down of the sun. *Astrophys. J.* **243**, 625–640 (1981).
26. Houghton, J. T., Meira Filho, L. G., Callander, B. A., Kattenberg, A. & Maskell, K. (eds) *Climate Change 1995: The Science of Climate Change* (Cambridge Univ. Press, Cambridge, 1996).
27. Degens, E. T. & Epstein, S. Relationship between ¹⁸O/¹⁶O ratios in coexisting carbonates, cherts and dolomites. *Am. Ass. Petrol. Geol. Bull.* **46**, 534–542 (1962).
28. Railsback, L. B. Influence of changing deep ocean circulation on the Phanerozoic oxygen isotopic record. *Geochim. Cosmochim. Acta* **54**, 1501–1509 (1990).
29. Knauth, L. P. & Epstein, S. Hydrogen and oxygen isotope ratios in nodular and bedded cherts. *Geochim. Cosmochim. Acta* **40**, 1095–1108 (1976).
30. Hays, P. D. & Grossman, E. L. Oxygen isotopes of meteoric calcite cements as indicators of continental paleoclimate. *Geology* **19**, 441–444 (1991).

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (Leibniz Prize and research grants to J.V.) and by the Natural Sciences and Engineering Research Council of Canada. International cooperation was facilitated by the Canadian Institute for Advanced Research (Toronto), with support from NORANDA and Hatch Investments Ltd. Y.G. and L.M.F. were supported by the Belgian National Foundation for Scientific Research (FNRS).

Correspondence and requests for materials should be addressed to J.V. (e-mail: veizer@science.utoronto.ca).

Evidence from Sardinian basalt geochemistry for recycling of plume heads into the Earth's mantle

D. Gasperini^{*,†}, J. Blichert-Toft^{*}, D. Bosch[‡], A. Del Moro[§], P. Macera[†], P. Télouk^{*} & F. Albarède^{*}

^{*} Ecole Normale Supérieure, 69364 Lyon Cedex 07, France

[†] Dipartimento di Scienze della Terra, Università degli Studi di Pisa, Via S. Maria, 53, 56126 Pisa, Italy

[‡] Université Montpellier 2, 34095 Montpellier Cedex 05, France

[§] Istituto di Geocronologia e Geochimica Isotopica, CNR, Via Alfieri, 1, 56010 Ghezzano-Pisa, Italy

Up to 10 per cent of the ocean floor consists of plateaux¹—regions of unusually thick oceanic crust thought to be formed by the heads of mantle plumes. Given the ubiquitous presence of recycled oceanic crust in the mantle source of hotspot basalts, it follows that plateau material should also be an important mantle constituent. Here we show that the geochemistry of the Pleistocene basalts from Logudoro, Sardinia, is compatible with the remelting of ancient ocean plateau material that has been recycled into the mantle. The Sr, Nd and Hf isotope compositions of these basalts do not show the signature of pelagic sediments. The basalts' low CaO/Al₂O₃ and Ce/Pb ratios, their unradiogenic ²⁰⁶Pb and ²⁰⁸Pb, and their Sr, Ba, Eu and Pb excesses indicate that their mantle source contains ancient gabbros formed initially by plagioclase accumulation, typical of plateau material. Also, the high Th/U ratios of the mantle source resemble those of plume magmas. Geochemically, the Logudoro basalts resemble those from Pitcairn Island, which contain the controversial EM-1 component that has been interpreted as arising from a mantle source sprinkled with remains of pelagic sediments^{2,3}. We argue, instead, that the EM-1 source from these two localities is essentially free of sedimentary material, the geochemical characteristics of these lavas being better explained by the presence of recycled oceanic plateaux. The storage of plume heads in the deep mantle through time offers a convenient explanation for the persistence of chemical and mineralogical layering in the mantle.

Variations of ^δ18O in Hawaiian basalts indicate that their mantle source contains material that was once exposed to alteration at low to medium temperatures. The positive correlation between ^δ18O and the ¹⁸⁷Os/¹⁸⁸Os ratio for the different Hawaiian volcanoes strongly suggests that the component altered at the lowest temperature (Koolau) has the largest time-integrated Re/Os ratio and therefore represents the upper layers of an ancient oceanic crust⁴. The other end-member of this correlation (Mauna Kea) was identified with the lower gabbroic oceanic crust, a conjecture strongly supported by the discovery of Sr anomalies in melt inclusions from Mauna Loa lavas⁵. In addition, the rather radiogenic Hf in the Koolau component and the curvature of the Hf–Pb isotope correlation for Hawaiian volcanoes require the presence of ancient pelagic sediments in the source of Hawaiian basalts⁶.

More components have been identified in the source of other hotspot basalts. The large U/Pb ratio ('HIMU'; that is, 'high μ ', where $\mu = ^{238}\text{U}/^{204}\text{Pb}$) characteristic of St Helena and some Polynesian basalts has been ascribed to the presence of ancient sea-floor basalts depleted in Pb and enriched in seawater U by hydrothermal activity³. Yet other hotspots, such as Pitcairn, contain a more controversial component (EM-1) that has some trace-element and isotope characteristics reminiscent of pelagic and/or metalliferous sediments^{2,3}. Alternatively, EM-1 has been compared with melts from the subcontinental lithospheric mantle (see discussion in ref. 7). Because of its nearly ubiquitous character in ocean island

basalts, EM-1 is central to the understanding of mantle component topology^{8,9}. Here we report an extensive set of geochemical data on Pleistocene basalts from Logudoro that represent an extreme case of EM-1 basalts.

The Logudoro volcanic fields are located in northwestern Sardinia, where the local thickness of the continental crust is 20–30 km (ref. 10). In order to assess whether the underlying continental crust affected the geochemistry of these volcanic rocks, a contemporaneous sample from the Tyrrhenian Sea, very similar to the Logudoro samples and collected during Leg 107 of the Ocean Drilling Project at Site 654¹¹, was also analysed. The geochemical characteristics of this sample do not differ from those of the Sardinian samples (see Supplementary Information). The isotopic compositions of Nd and Hf of the Logudoro volcanics and, to a lesser extent, their Sr isotope compositions, plot on the enriched extension of the normal mantle array (Fig. 1). Hf is not particularly radiogenic with respect to Nd

and, unlike the situation for Hawaiian basalts⁶, does not plot towards the field of pelagic sediments above the Hf–Nd array. The Nb/U and Ce/Pb ratios have been widely used to discriminate between oceanic basalts and those from the continental crust^{12,13}. As such, they have become indicators of continental-crust influence, either in the form of a source component or acquired by shallow-level contamination. The average Nb/U ratio of Logudoro volcanics (48) is typical of uncontaminated oceanic basalts (47 ± 10), whereas the average Ce/Pb ratio is lower (15 with respect to 25 ± 5). The Ce/Pb ratios seem to be weakly correlated with $^{206}\text{Pb}/^{204}\text{Pb}$ (correlation coefficient of +0.5), whereas a negative correlation would be expected from contamination by typical continental material ($\text{Ce/Pb} < 5$, $^{206}\text{Pb}/^{204}\text{Pb} > 18$). Lower continental crust with unradiogenic Pb and low Ce/Pb would be a potentially acceptable contaminant, but the similarity between the Pb isotope compositions of the ODP Leg 107 sample and the samples erupted

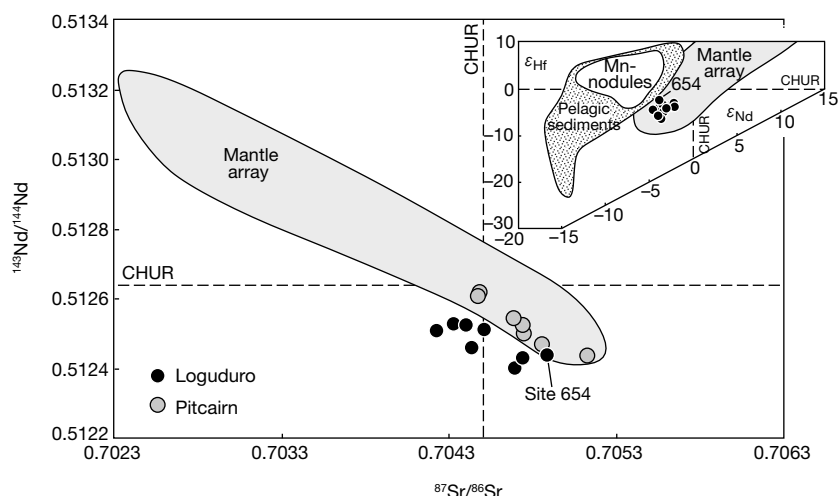


Figure 1 Plot of $^{143}\text{Nd}/^{144}\text{Nd}$ versus $^{87}\text{Sr}/^{86}\text{Sr}$ for basalts from Logudoro (Sardinia), Site 654 (sample 654-9-CC-00-04) and Pitcairn. Inset, ϵ_{Hf} versus ϵ_{Nd} for basalts from Logudoro and Site 654 (sample 654-9-CC-00-04). The mantle array is contoured from a

large amount of published and unpublished data, too numerous to reference here, but available from J.B.-T. Fields for Mn nodules and pelagic sediments are from refs 26 and 27, respectively. CHUR, chondritic uniform reservoir.

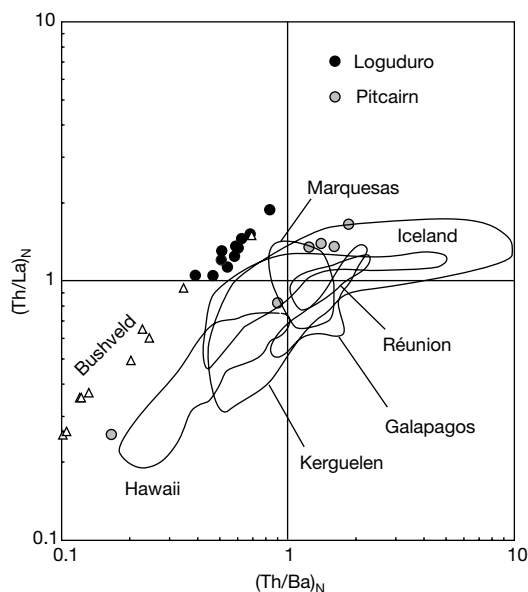


Figure 2 Plot of normalized Th/La versus normalized Th/Ba (ref. 14) in basalts from Logudoro (Sardinia) and Pitcairn¹⁶. The fields of other ocean island basalts are also shown, and Bushveld gabbros¹⁵ are shown for reference. The ratios are normalized to the primitive mantle values of ref. 28. The fields were drawn from a compilation of values obtained from the GEOROC database (<http://georoc.mpch-mainz.gwdg.de>). During

magmatic processes, Ba fractionates into plagioclase. Excess Ba therefore signals the presence of a cumulative plagioclase component (metagabbro) in the source of the Logudoro basalts. With plagioclase being on the liquidus of basalts at pressures of less than 1 GPa, this observation rules out the subcontinental lithospheric mantle as the source of these basalts.

in Sardinia does not support extensive incorporation of crustal Pb. A petrogenetic model that does not alter the source geochemical characteristics of the Logudoro basalts through contamination by crust, including the presence of a terrigenous component in the mantle source, must therefore be found.

The Logudoro lavas are essentially aphyric alkali basalts and trachybasalts with olivine and plagioclase microphenocrysts. Major-element features include Mg# between 46 and 64 and high Al_2O_3 contents (15–17%; see Table 1 in Supplementary Information). The remarkably low $\text{CaO}/\text{Al}_2\text{O}_3$ ratios (0.4) of these rocks are independent of the MgO contents. Such a relationship excludes substantial pyroxene fractionation during the petrogenesis of the lavas, and signals the abundance of a plagioclase-rich component in the source. We contend that this component consists of metagabbros, that is, ancient plagioclase-rich cumulates entrained to depth by ancient subduction and processed in the deep mantle. The unusual presence of an abundant plagioclase component in the source is also indicated by the excesses of four elements, Ba, Sr, Eu and Pb, that preferentially partition into this mineral during magmatic processes. Ba enrichment is particularly visible in a plot of Th/La versus Th/Ba¹⁴ (Fig. 2), where the Logudoro volcanics fall on the extension of the Bushveld anorthosites¹⁵. The rather large correlation coefficient (+0.63) between Sr/Eu^* and Eu/Eu^* (Fig. 3)—where Eu^* is the Eu concentration estimated by interpolating between the neighbouring trivalent elements Sm and Gd—indicates that the Sr excess correlates with the positive Eu anomaly and therefore suggests control of the source composition by plagioclase fractionation. The $^{206}\text{Pb}/^{204}\text{Pb}$ and $^{207}\text{Pb}/^{204}\text{Pb}$ ratios plot to the left of the geochron (Fig. 4b). The low time-integrated U/Pb and Th/Pb ratios inferred from the unradiogenic Pb isotope compositions concur with the low Ce/Pb ratios of the rocks, and confirm that the mantle source of the Logudoro basalts was dominated by a plagioclase-bearing component for a long period of time.

The Pb isotope compositions are among the least radiogenic of all modern terrestrial basalts (Fig. 4). The combination of unradiogenic ^{206}Pb and normal ^{208}Pb further requires that the source had a high time-integrated Th/U ratio ($\kappa = 4.1$), a feature found only in

EM-1 basalts such as those from Pitcairn and a few Polynesian islands^{3,16}. Pitcairn and Logudoro basalts otherwise have rather similar Sr, Nd and Hf isotope compositions (Fig. 1) and Ce/Pb and Nb/U ratios. Pitcairn has more radiogenic ^{208}Pb (Fig. 4a) and less radiogenic ^{207}Pb (Fig. 4b). The relative distributions of the most incompatible elements in the two series are similar. All the incompatible elements, except Ba and Sr, are depleted in Pitcairn relative to Logudoro basalts. We suggest that the source of these two series initially contained various gabbros and dolerites that were related to each other by low-pressure plagioclase fractionation. From the relative abundances of Sr, Ba and Pb in the two series, the protolith appears to be more cumulative and less differentiated for Logudoro than for Pitcairn basalts. The basaltic samples from the Walvis ridge and, to a lesser extent, Tristan da Cunha, resemble those from Pitcairn.

Although the EM-1 basalts from Logudoro and Pitcairn share evidence with some Hawaiian lavas for plagioclase control of their protolith⁵, differences in their isotopic compositions signal that the EM-1 sources formed in another manner. The presence of abundant plagioclase on the liquidus of the igneous rocks that eventually contributed to the source of EM-1 basalts is a strong argument against recycled subcontinental lithospheric mantle. As plagioclase does not reach saturation in basaltic liquids at pressures in excess of 1 GPa (see, for example, ref. 17), it cannot appear at mantle depths below the continental Moho (35–45 km). In contrast with Hawaiian volcanics, which show a relatively strong time-integrated depletion in highly incompatible elements (radiogenic Nd and Hf, unradiogenic Sr), the Logudoro samples, just as those from Pitcairn, show a slight time-integrated enrichment (Fig. 1). The high Th/U ratio of

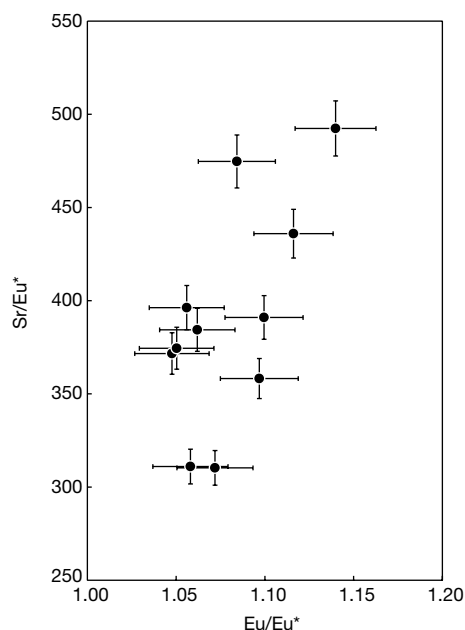


Figure 3 Plot of Sr/Eu^* versus Eu/Eu^* in the basalts from Logudoro. Eu^* is the Eu concentration derived by interpolation between Sm and Gd. The high value of the correlation coefficient (0.63) indicates that Eu and Sr excesses have a common cause, namely plagioclase accumulation in the gabbro that was the precursor of the mantle source of the Logudoro volcanics.

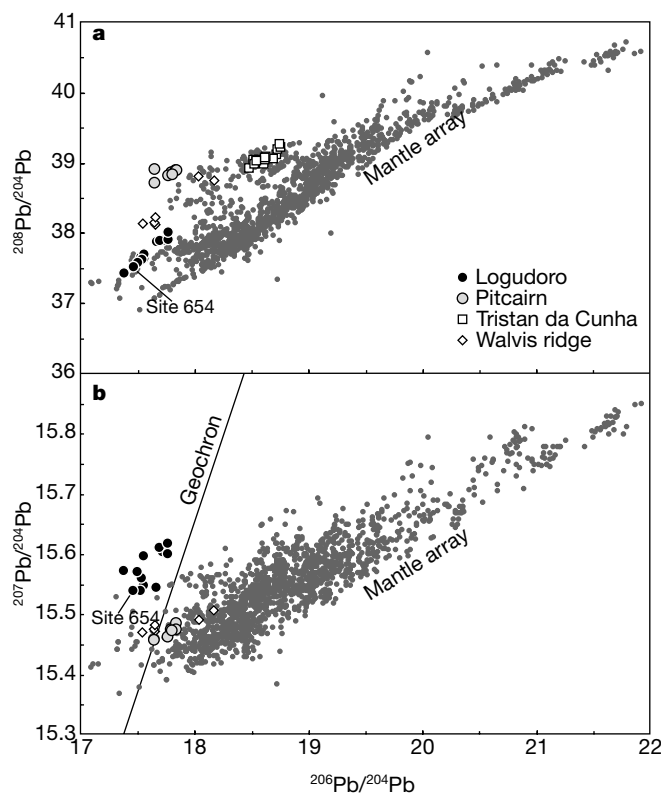


Figure 4 Lead isotope compositions of basalts from Logudoro (Sardinia), Site 654 (sample 654-9-CC-00-04), Pitcairn, Walvis ridge and Tristan da Cunha. **a**, $^{208}\text{Pb}/^{204}\text{Pb}$ versus $^{206}\text{Pb}/^{204}\text{Pb}$. **b**, $^{207}\text{Pb}/^{204}\text{Pb}$ versus $^{206}\text{Pb}/^{204}\text{Pb}$. The mantle array is contoured from hundreds of data points from the PETDB database for MORBs (<http://petdb.ldeo.columbia.edu/petdb>) and the GEOROC database for ocean island basalts. Data for Walvis ridge and Tristan da Cunha are from refs 29 and 30, respectively. The age adopted for the geochron is 4.55 Gyr. The unradiogenic character of ^{206}Pb and ^{208}Pb in the Logudoro basalts shows that the source had low U/Pb and Th/Pb ratios, a further indication that it formed by plagioclase accumulation.

the mantle source of Logudoro and Pitcairn basalts is another distinctive feature that does not fit a mantle source dominated by mid-ocean-ridge basalt (MORB).

It is clear that normal oceanic crust (N-MORB), either ancient or recent, is an unsuitable mantle source for the Logudoro and Pitcairn basalts. Because the above discussion has concluded that ancient gabbros and dolerites are components of the source, basaltic series with only a mild enrichment of the most incompatible elements should be sought. Plateau basalts, and, in particular, oceanic plateaux—obviously abundant and formed continuously throughout the Earth's history, such as those from Ontong-Java or the Caribbean¹⁸—form a possible mantle source component that can account almost ideally for the characteristics of EM-1 basalts. Rare-earth-element patterns that range from nearly flat to slightly enriched, the lack of a Nb anomaly, and low-pressure conditions of differentiation giving rise to shallow cumulates rich in plagioclase, are all conditions that should bring both the trace-element distributions and long-term isotopic patterns to resemble EM-1 basalt mantle sources.

Another potential mantle source of EM-1 basalts could be anomalous, geochemically enriched, oceanic crust. This material, referred to as E-MORB, does not represent a well defined component but is rather the enriched 'tail' of the normal population of ridge basalts: ultra-depleted MORB represents the other 'tail' of the distribution. E-MORB tends to form shallow ridges¹⁹ and grades into plume material. The most compelling argument in favour of a plume source, and against any type of MORB source, for the EM-1 basalts is their time-integrated Th/U (κ) value derived from their Pb isotope compositions; these κ values vary between 4.0 and 4.5. In comparison, all fresh MORB glasses form a unimodal population with an average Th/U value of 2.5, in agreement with previous estimates²⁰ and with ²³⁰Th isotope geochemistry²¹, whereas ocean island basalts have substantially higher values, such as observed for Réunion (4.0), Galapagos (3.7), Hawaii (3.0) and Iceland (3.3). Any material that is part of the EM-1 source falls outside the MORB range, whereas normal plume material displays much more suitable Th/U ratios.

In contrast to oceanic crust, large buoyant oceanic aseismic ridges and plateaux are not usually subducted²². Such ridges and plateaux are now considered as dominant sources of juvenile material accreting to the continents^{1,23}. But small or thin plateaux, associated with minor plume heads, as well as small fragments of larger plateaux attached to large heavy sinking plates, have less buoyancy and may be entrained by subduction into the mantle where they subsequently evolve into a potential mantle component for basaltic melts in much the same way as does normal oceanic crust. As the different altered basaltic sections of the oceanic crust, including its overlying pelagic sediments, have been identified as normal components of the source of hotspot basalts, it is not surprising that plateaux, which are just another common element of oceanic lithospheric plates, are also now found to be a component of oceanic basalt. The 'graveyard' of the plates must accommodate all those plume heads formed during the entire history of the Earth that escaped incorporation into the continental crust. We propose here that Logudoro and Pitcairn basalts are melts from precisely the mantle component that represents recycled oceanic plateaux.

The geochemical characteristics of the EM-1 reservoir sampled by Logudoro and Pitcairn basalts can also mitigate the apparent imbalance between terrestrial heat flow and heat production. This imbalance necessitates the presence of U, Th and K in inaccessible deep mantle layers in concentrations higher than those of the depleted upper mantle²⁴—and even higher than those of the subducted oceanic plates imaged by high-resolution mantle tomography²⁵. As discussed above, the EM-1 reservoir is unlike oceanic crust, but is consistent with mildly enriched material formed from plume head magmas similar to those of oceanic plateaux. The different shape and chemistry of plume heads compensate for their apparently small proportion with respect to

that of oceanic crust. The crust beneath the Ontong-Java plateau is 4–6 times as thick as normal oceanic crust, and U and Th are 3–5 times more concentrated in its basalts than in normal MORB¹⁸. The preferential storage of plume material in the deep mantle may be one of the factors that help sustain the chemical and mineralogical zoning that has been inferred from seismic tomography and heat-flow measurements. □

Received 22 February; accepted 21 September 2000.

- Ben-Avraham, Z., Nur, A., Jones, A. & Cox, D. Continental accretion: from oceanic plateaux to allochthonous terranes. *Science* **213**, 47–54 (1981).
- Weaver, B. L. The origin of ocean island basalt end-member compositions: trace element and isotopic constraints. *Earth Planet. Sci. Lett.* **104**, 381–397 (1991).
- Chauvel, C., Hofmann, A. W. & Vidal, P. HIMU-EM: the French Polynesian connection. *Earth Planet. Sci. Lett.* **110**, 99–119 (1992).
- Lassiter, J. C. & Hauri, E. H. Osmium-isotope variations in Hawaiian lavas: evidence for recycled oceanic lithosphere in the Hawaiian plume. *Earth Planet. Sci. Lett.* **164**, 483–496 (1998).
- Sobolev, A. V., Hofmann, A. W. & Nikogosian, I. K. Recycled oceanic crust observed in 'ghost plagioclase' within the source of Mauna Loa lavas. *Nature* **404**, 986–990 (2000).
- Blichert-Toft, J., Frey, F. A. & Albarède, F. Hf isotope evidence for pelagic sediments in the source of Hawaiian basalts. *Science* **285**, 879–882 (1999).
- Hofmann, A. W. Mantle geochemistry: the message from oceanic volcanism. *Nature* **385**, 219–229 (1997).
- Hart, S. R., Hauri, E. H., Oschmann, L. A. & Whitehead, J. A. Mantle plumes and entrainment: isotopic evidence. *Science* **256**, 517–520 (1992).
- Hanan, B. B. & Graham, D. W. Lead and helium isotope evidence from oceanic basalts for a common deep source of mantle plumes. *Science* **272**, 991–995 (1996).
- Morra, V., Secchi, F. A. G., Melluso, L. & Franciosi, L. High-Mg subduction related Tertiary basalts in Sardinia, Italy. *Lithos* **40**, 69–91 (1997).
- Beccaluva, L. *et al.* Geochemistry and mineralogy of volcanic rocks from ODP sites 650, 651, 655 and 654 in Tyrrhenian Sea. *Proc. ODP Sci. Res.* **107**, 49–74 (1990).
- Hofmann, A. W., Jochum, K. P., Seufert, M. & White, W. M. Nb and Pb in oceanic basalts: new constraints on mantle evolution. *Earth Planet. Sci. Lett.* **79**, 33–45 (1986).
- Sims, K. W. W. & DePaolo, D. J. Inferences about mantle magma sources from incompatible element concentration ratios in oceanic ratios. *Geochim. Cosmochim. Acta* **61**, 765–784 (1997).
- Hofmann, A. W. & Jochum, K. P. Source characteristics derived from very incompatible trace elements in Mauna Loa and Mauna Kea basalts, Hawaii scientific drilling project. *J. Geophys. Res.* **101**, 11831–11839 (1996).
- Arndt, N. T., Wilson, A. H., Lee, C. A. & Ohnenstetter, M. In *Proc. Bushveld Workshop 6* (Lokopdam, South Africa, 1999).
- Woodhead, J. D. & McCulloch, M. T. Ancient seafloor signals in Pitcairn Island lavas and evidence for large amplitude, small length-scale mantle heterogeneities. *Earth Planet. Sci. Lett.* **94**, 257–273 (1989).
- Ringwood, A. E. *Composition and Petrology of the Earth's Mantle* (McGraw-Hill, New York, 1975).
- Neal, C. R., Mahoney, J. J., Kroenke, L. W., Duncan, R. A. & Petterson, M. G. (eds) *The Ontong-Java plateau* (American Geophysical Union, Washington DC, 1997).
- Klein, E. M. & Langmuir, C. H. Global correlations of ocean ridge basalt chemistry with axial depth and crustal thickness. *J. Geophys. Res.* **92**, 8089–8115 (1987).
- White, W. M. ²³⁸U/²⁰⁴Pb in MORB and open system evolution of the depleted mantle. *Earth Planet. Sci. Lett.* **115**, 211–226 (1993).
- Allègre, C. J. & Condomines, M. Basalt genesis and mantle structure studied through Th–isotopic geochemistry. *Nature* **299**, 21–24 (1982).
- Vogt, P. R., Lowrie, A. & Bracey, D. R. Subduction of aseismic oceanic ridges: effects on shape, seismicity and other characteristics of consuming plate boundaries. *Geol. Soc. Am. Spec. Pap.* **172**, 1–59 (1976).
- Abouchami, W., Boher, M., Michard, A. & Albarède, F. A major 2.1 Ga event of mafic magmatism in West Africa on early stage of crustal accretion. *J. Geophys. Res.* **95**, 17605–17629 (1990).
- Kellogg, L. H., Hager, B. H. & van der Hilst, R. D. Compositional stratification in the deep mantle. *Science* **283**, 1881–1884 (1999).
- van der Hilst, R. S., Widiyantoro, E. R. & Engdahl, T. Evidence for deep mantle circulation from global tomography. *Nature* **386**, 578–584 (1997).
- Albarède, F., Simonetti, A., Vervoort, J. D., Blichert-Toft, J. & Abouchami, W. A Hf–Nd isotopic correlation in ferromanganese nodules. *Geophys. Res. Lett.* **20**, 3895–3898 (1998).
- Vervoort, J. D., Patchett, P. J., Blichert-Toft, J. & Albarède, F. Relationships between Lu–Hf and Sm–Nd isotopic systems in the global sedimentary system. *Earth Planet. Sci. Lett.* **168**, 79–99 (1999).
- McDonough, W. F. In *The Encyclopedia of Geochemistry* (eds Marshall, C. P. & Fairbridge, R. F.) 151–156 (Kluwer, Amsterdam, 1999).
- Richardson, S. H., Erlank, A. J., Duncan, A. R. & Reid, R. L. Correlated Nd, Sr and Pb isotope variation in Walvis Ridge basalts and implications for the evolution of their mantle source. *Earth Planet. Sci. Lett.* **59**, 327–342 (1982).
- Newsom, H. E., White, W. M., Jochum, K. P. & Hofmann, A. W. Siderophile and chalcophile element abundances in oceanic basalts, Pb isotope evolution and growth of the Earth's core. *Earth Planet. Sci. Lett.* **80**, 299–313 (1986).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank M. Murtas and A. Cosmelli for help with sample collection, and F. Frey, B. Hanan and A. Hofmann for comments on the manuscript. This work was supported by the Institut National des Sciences de l'Univers, Cofinanziamento MURST and the University of Pisa.

Correspondence and requests for materials should be addressed to J.B.-T. (e-mail: jblicher@ens-lyon.fr).

The smallest known non-avian theropod dinosaur

Xing Xu*, Zhonghe Zhou* & Xiaolin Wang*†

* The Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, PO Box 643, Beijing 100044, People's Republic of China

† College of Earth Science and Resources, China University of Geosciences, Beijing 100083, People's Republic of China

Non-avian dinosaurs are mostly medium to large-sized animals, and to date all known mature specimens are larger than the most primitive bird, *Archaeopteryx*¹. Here we report on a new dromaeosaurid dinosaur, *Microraptor zhaoianus* gen. et sp. nov., from the Early Cretaceous Jiufotang Formation of Liaoning, China². This is the first mature non-avian dinosaur to be found that is smaller than *Archaeopteryx*¹, and it eliminates the size disparity between the earliest birds and their closest non-avian theropod relatives. The more bird-like teeth, the *Rahonavis*-like ischium and the small number of caudal vertebrae of *Microraptor* are unique among dromaeosaurids and improve our understanding of the morphological transition to birds. The nearly completely articulated foot shows features, such as distally positioned digit I, slender and recurved pedal claws, and elongated penultimate phalanges, that are comparable to those of arboreal birds^{3–6}. The discovery of these in non-avian theropods provides new insights for studying the palaeoecology of some bird-like theropod dinosaurs.

Theropoda Marsh 1881

Maniraptora Gauthier 1986

Dromaeosauridae Matthew & Brown 1922

Microraptor zhaoianus gen. et sp. nov.

Etymology. The generic name refers to the small size of this new dromaeosaurid dinosaur; the specific name is in honour of Zhao Xijin, a distinguished dinosaurologist who introduced the first author to the field of vertebrate paleontology.

Holotype. IVPP (The Institute of Vertebrate Paleontology and Paleoanthropology) V 12330, a partial skeleton preserved in main (Fig. 1) and counter slabs.

Locality and horizon. Xiasanjiazi, Chaoyang County, western Liaoning; Jiufotang Formation (Early Cretaceous²), underlain by the famous Yixian Formation which has yielded more than 1,000 specimens of early birds and feathered dinosaurs⁷.

Diagnosis. Distinguishable from other dromaeosaurids in anterior serrations are absent on all of the teeth; posterior teeth have a basal constriction between crown and root; middle caudals are about 3–4 times as long as the anterior dorsals; accessory crest of femur at the base of the lesser trochanter; the tail has less than 26 vertebrae; and it has strongly recurved and slender pedal unguals with a prominent flexor tubercle.

Description. The holotype of *M. zhaoianus* is an articulated skeleton missing the middle portion of the body (Figs 1 and 2). The premaxilla is similar to that of *Archaeopteryx*⁸ and *Sinornithosaurus*⁹ in that it has a sloping anterior margin. As in *Archaeopteryx* and troodontids¹⁰, the maxilla contributes to the border of the external naris (Fig. 2b). There are at least 15 dentary teeth, which are closely packed as in troodontids¹¹. The anterior teeth, including premaxillary teeth, are more recurved and laterally compressed than the posterior teeth as in most theropods, but they lack serrations on both the anterior and the posterior carinae as seen in some non-avian theropods and birds (Fig. 2c). The posterior teeth have posterior serrations; however, they are bird-like in having a less compressed crown and a constriction between the root and crown (Fig. 2d). These 'waisted' teeth are reported here for the first time in a dromaeosaurid, although among non-avian theropods¹² they have already been reported in therizinosaurs, troodontids and ornithomimosaurs. The heterodont dentition provides information on the transition from the non-avian theropod type of dentition to that of the typical avian type. It seems that the reduction of serrations begins in the anterior teeth and with the anterior margins of the teeth, however, the basal constriction of the tooth crown first appears in the posterior teeth.

The body is extremely short, with an estimated trunk length of about 47 mm (this length is inferred from the presumption that *Microraptor* had 13 dorsal bones and that its cervicodorsals are relatively long as compared with the remaining dorsals, as in other dromaeosaurids¹³). Five fused sacral bones are preserved and the anterior three are transversely enlarged, although the sacra of the holotype are slightly crushed and flattened (Fig. 3a). The tail is almost complete and articulated with the sacrum. There are 24 or 25

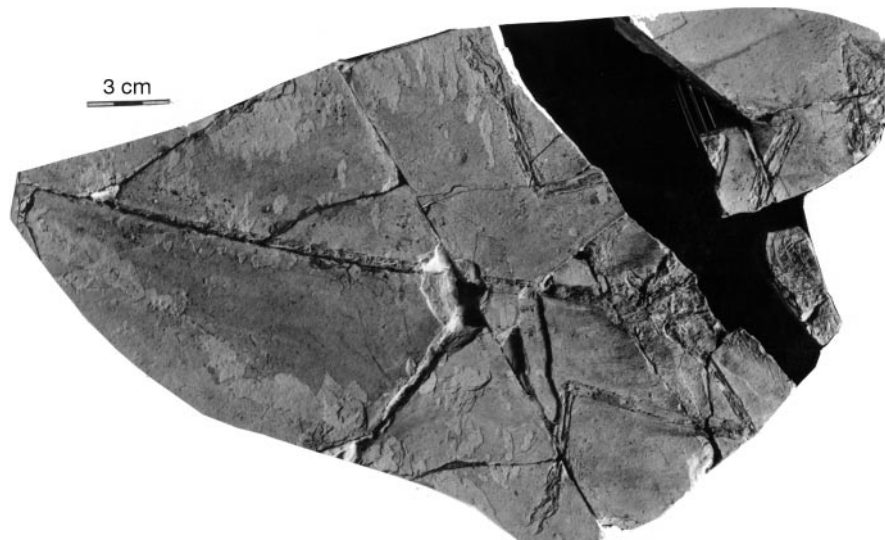


Figure 1 Main slab (V 12330, holotype) containing *M. zhaoianus*. V 12330 includes a partial skull, complete mandible, right radius, right ulna, partial right manus, incomplete right hindlimb and nearly complete left hindlimb, dorsals, sacra, nearly complete tail, and ribs, and nearly complete pelvic girdle. Examination of the preservation and

proportions of the elements indicate that the material represents a single individual. The tail portion of the single known specimen of '*Archaeopteryx*' forms the counterpart of the tail of V 12330 and therefore represents part of the holotype of *Microraptor zhaoianus*.

caudals, which is close to the caudal number of *Archaeopteryx*¹⁴. The articulated tail has elongate rod-like extensions of the prezygapophyses and chevrons characteristic of dromaeosaurid dinosaurs^{13,15}. These rod-like extensions reach almost to the sacrum (Fig. 3a). The caudals are significantly elongated. Gastralria and what appear to be sternal ribs and uncinat processes are preserved.

The ulna is bowed posteriorly and when compared with the femur⁹ it is proportionately shorter than in other dromaeosaurids; the radius is much thinner than the ulna. There are at least three carpals and an additional one may be present (Fig. 3b), a pattern that is similar to that of *Archaeopteryx*. The semilunate carpal is relatively small as in *Sinornithosaurus* (IVPP, V 12811) and in *Archaeopteryx*^{16,17}, and has more contact with the proximal end of

metacarpal II than with metacarpal I (Fig. 3b).

The pubis is probably retroverted as in other dromaeosaurids and birds (Fig. 3a), and its symphysis is about 49% the length of the pubis. The ischium is plate-like and less than half the length of the pubis; it has a posterior process as in *Sinornithosaurus*⁹, *Unenlagia*, *Archaeopteryx* and *Rahonavis*¹⁸. A large distally positioned obturator process is present as in *Rahonavis*¹⁸ and *Sinornithosaurus* (IVPP, V 12811).

The femur of *M. zhaoianus* has an accessory crest pendent at the base of the lesser trochanter, as in the basal oviraptorosaurian *Microvenator*¹⁹. The tibia is about 128% the length of the femur. The fibula is slender and reaches the calcaneum. Pedal digit I is preserved on the posterior surface of metatarsals II–IV in both of the hindlimbs and has a more distal position than in other known non-avian theropods. Metatarsals II, III and IV are subequal in length. The proximal end of metatarsal III is significantly compressed in posterior view, which suggests an arctometatarsalian condition²⁰. Metatarsal IV is much more robust than metatarsals II and III, which is also the case in troodontids. Metatarsal V is long and bowed laterally and its mid-shaft is expanded. Metatarsal IV bears a pronounced flange on the posteromedial surface as in other dromaeosaurids^{9,21}. The second pedal digit is a specialized raptorial tool¹³ as in other dromaeosaurids, troodontids and *Rahonavis*. Like other dromaeosaurids²², the ungual is more than one and a half times longer than phalanx II-1. All of the pedal unguals are sharp and strongly recurved, similar to those of trunk-climbing birds^{4–6}. The horny sheath is well preserved, with a columnar dorsal margin

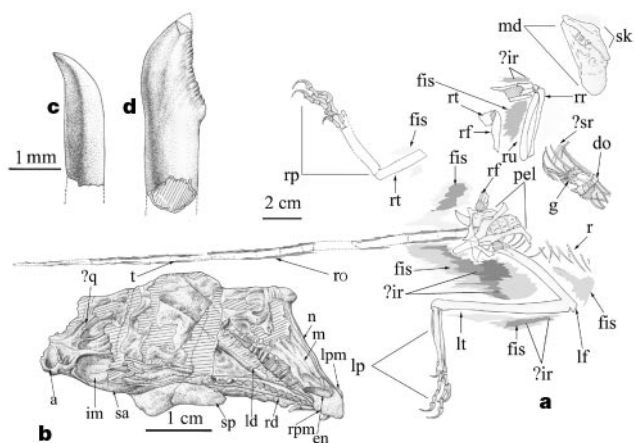


Figure 2 Outline of skeleton and identification of the skull with lower jaws and an anterior and posterior dentary tooth of *M. zhaoianus* (V 12330). **a**, Outline of the skeleton (broken lines indicate bones preserved only as impressions; stippled line indicates damaged bone). **b**, Skull with lower jaws. **c**, An anterior dentary tooth in medial view. **d**, A posterior dentary tooth in lateral view. a, articular; do, dorsal vertebra; en, external naris; fis, feather-like integumentary structures; g, gastralria; im, internal mandibular fossa; ir, impression of rachis; lf, left femur; lp, left pes; lpm, left premaxilla; lt, left tibia; m, maxilla; md, mandible; n, nasal; pel, pelvis; q, quadrate; r, rib; rd, right dentary; rf, right femur; ro, rod-like extensions; rp, right pes; rpm, right premaxilla; rr, right radius; rt, right tibia; ru, right ulna; sp, splenial; sa, surangular; sk, skull; sr, sternal rib; and t, tail. Question mark indicates uncertain identification.

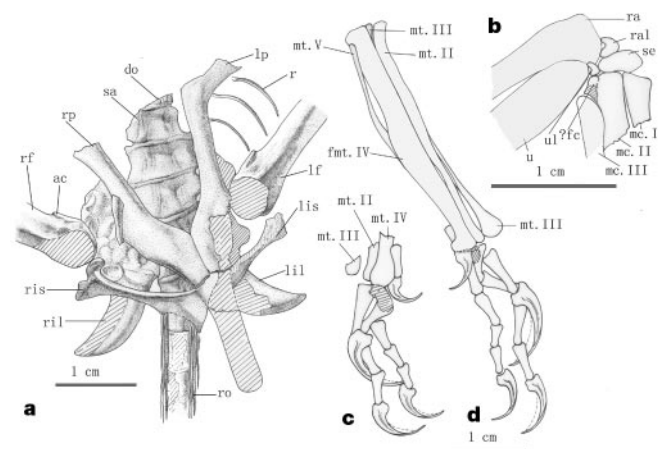


Figure 3 Identification of pelvic girdle, right forelimb and pes of *M. zhaoianus* (V 12330). **a**, Pelvic girdle in ventral view. **b**, Right forelimb in ventral view. **c**, Right pes in posterior view. **d**, Left pes in posterior view. Additional abbreviations: ac, accessory crest; fc, fourth carpal; fnt. IV, flange on metatarsal IV; lil, left ilium; lis, left ischium; lp, left pubis; mc. I–III, metacarpals I–III; mt. II–V, metatarsals II–V; ra, radius; ral, radiale; ril, right ilium; ris, right ischium; rp, right pubis; sa, sacrum; se, semilunate distal carpal; and u, ulna. See also Fig. 3 for abbreviations. Question mark indicates uncertain identification.

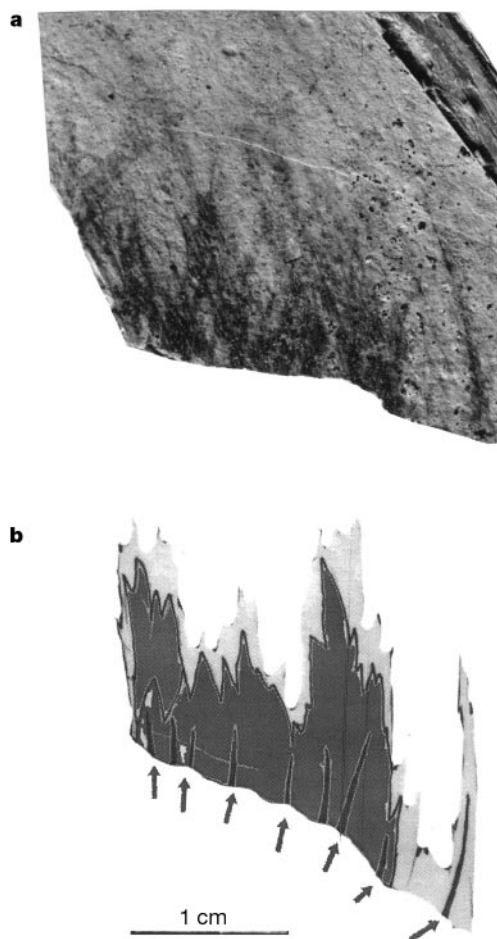


Figure 4 Integument of *M. zhaoianus* (V 12330). Photograph (**a**), and line drawing (**b**), of parts of the integument preserved adjacent to the femur. Shaded areas represent the carbonized integument, and arrows point to the median depression that was probably made by the rachis.

and a thin ventral margin.

Large patches of integuments are preserved *in situ* around the skeleton (Figs 1 and 2a), and the pattern of preservation is similar to that of early birds from the same locality (IVPP, V 12325). Integumentary structures are best preserved near the femur, where they run almost perpendicular to the bone (Fig. 4). They are long (average 25–30 mm), narrow, and have a feather-like contour, whereas those along the tibia and in the hip area are shorter. Some impressions of the integuments contain a structure similar to that of a rachis (Fig. 4b), suggesting that true feathers may have been present among dromaeosaurids.

Discussion. We investigated the phylogenetic relationship of *Microraptor* by means of a cladistic analysis of a data set comprising 13 taxa and 89 characters (see Supplementary Information). Our analysis suggests that *Microraptor* is a basal dromaeosaurid (Fig. 5). Notable in this new dromaeosaurid is the presence of some salient similarities to those of troodontids, such as troodontid-like posterior teeth and arctometatarsalian metatarsals. The discovery of *Microraptor* completes some morphological gaps between dromaeosaurid dinosaurs and birds; for example, it has more bird-like teeth than do other dromaeosaurids, a *Rahonavis*-like ischium, enlarged sacra and fewer caudal vertebrae. These bird-like characteristics were more widespread and retained in the lineage leading to the birds and in the basal dromaeosaur *Microraptor*, but were lost (by means of reversal to a more primitive state) in more derived dromaeosaurids.

There are a few notable features of *Microraptor*, one of which is its small size. The holotype of *Microraptor* is a mature animal, as indicated by the fusion of the sacral vertebrae, the pubic symphysis and the partial fusion of the last dorsal to the synsacrum. The fine serrations on the tooth crown, the well-developed accessory crest on the femur and the well-ossified cortical bone provide further

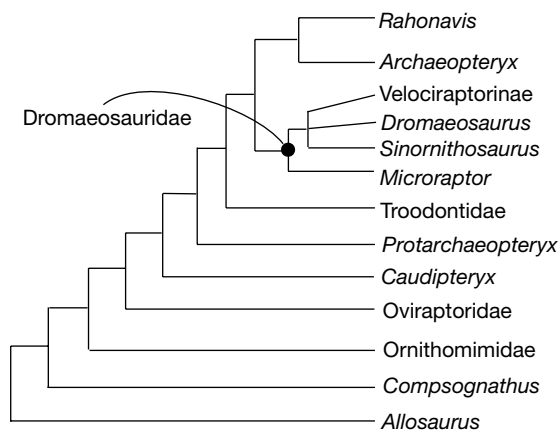


Figure 5 Phylogenetic relationships of *M. zhaoianus*. Shown is a strict consensus tree of the two most parsimonious trees (tree length = 170; consistency index = 0.594; retention index = 0.667). Characters 1–3 are ordered. The tree was produced using the branch-and-bound option of PAUP. Synapomorphies for each node were determined by accelerated transformation (ACCTRAN). *M. zhaoianus* was referred to Dromaeosauridae on the basis of the following unique dromaeosaurid features: ossified caudal rods extending the length of prezygapophyses and chevrons present (80.1); a longitudinal ridge along the posteromedial surface of metatarsal IV present (87.1); second pedal claw apparently larger than the pedal phalanx II-1 (89.1). Additional synapomorphies: 2.1, 3.1, 10.1, 66.1, 68.1, 70.1, 74.1, 76.1, 79.1, 63.1, 64.1, 65.1, 67.1, 69.1 and 75.1.

M. zhaoianus is less derived than other dromaeosaurids and lacks the following synapomorphies: denticles present on anterior and posterior carinae of teeth (1.0); all teeth laterally compressed and recurved (2.0); constriction between crown and root absent on all teeth (3.0); maxilla contribution to naris absent (4.0); number of caudal vertebrae more than 30 (20.0); femoral neck present (53.0); tibia less than 115% length of femur (53.0); and internal mandibular fenestra triangular and relatively large (78.1). The monophyletic Dromaeosauridae tree did not collapse until four more steps had been added by re-analysing the data matrix and sequentially adding steps.

evidence of its maturity^{19,23}. Additional evidence includes its relatively small skull (skull/femur length ratio is 0.85) and sacrum (sacrum/femur length ratio is 0.36, compared with 0.44 in *Sinornithosaurus*; in *Archaeopteryx* the larger individuals have a smaller sacrum/femur length ratio than in *Microraptor*).

Although a number of non-avian theropods of small size have been reported previously, all are larger than the most primitive bird *Archaeopteryx*¹ (see Supplementary Information). The trunk of *Microraptor* is shorter than in all but the smallest individual of *Archaeopteryx*²⁴; the sacrum is as long as in the smallest individual of *Archaeopteryx*; and the femur is as long as that of the Berlin specimen²⁴, which is the third smallest specimen of *Archaeopteryx* found (see Supplementary Information). These comparisons indicate that *Microraptor* has the smallest adult size of any non-avian dinosaur found to date. Decrease of body size is an evolutionary trend among coelurosaurs¹ and had an important role in the origin of bird flight²⁵. The small size of *Microraptor* suggests that non-avian theropods were pre-adapted for the origin of flight. In contrast, a trend of increasing body size is evident among dromaeosaurids, and is also found in many other coelurosaurian lineages including tyrannosaurids²⁶, therizinosaurs²⁷ and oviraptorosaurs¹.

Some pedal features of *Microraptor* are consistent with an arboreal habit. Pedal digit I is relatively distal in position, as indicated by its first phalanx extending distally almost to the distal end of metatarsal III. This is similar to the situation in *Archaeopteryx* (see ref. 24), *Confuciusornis* (IVPP, V 11370) and arboreal birds, but is different from that of most non-avian theropods and ground birds in which pedal digit I is more proximally positioned. All pedal unguals of *Microraptor* are slender, with a curvature of more than 155° (calculated by the method of ref. 6) and a prominent flexor tubercle (Fig. 3c, d). They are different from all known non-avian theropods, but comparable to those of climbing birds^{4–6} (see Supplementary Information). The penultimate pedal phalanges of *Microraptor* and, to a lesser degree, other dromaeosaurids are elongated relative to those of most cursorial theropods (see Supplementary Information). The ratios of pedal phalanx II-2/II-1 are larger than 1 in dromaeosaurids and early birds but smaller than 1 in other non-avian maniraptorans. The elongation of the distal pedal phalanges has been suggested to be indicative of an arboreal habit^{25,28}. The available evidence indicates that *Microraptor* may have been able to climb trees, although this dinosaur is apparently derived from a cursorial ancestor as it retained an arctometatarsalian foot.

Non-avian theropods are usually regarded as being terrestrial cursors²⁹. The discovery of *Microraptor* provides the first evidence to suggest that some non-avian theropods had arboreal habits, indicating that the palaeoecology of non-avian theropods may be more diverse than previously thought. The evidence for arboreality in non-avian theropods is an important discovery as it could be used to test the 'tree-down' hypothesis for the origin of bird flight³⁰. The theropod ancestors of birds may have passed through an arboreal phase, although more evidence is needed to confirm this hypothesis. □

Received 24 March; accepted 29 September 2000.

- Sereno, P. C. The evolution of dinosaurs. *Science* **284**, 2137–2147 (1999).
- Jin, F. New advances in the late Mesozoic stratigraphic research of western Liaoning, China. *Vertebrata Palasiatica* **34**, 102–122 (1996).
- Yaden, D. W. In *The Beginnings of Birds* (eds Hecht, M. K., Ostrom, J. H., Viohl, G. & Wellnhofer, P.) 91–97 (Freunde des Jura-Museums Eichstatt, Eichstatt, 1985).
- Yalden, D. W. Climbing *Archaeopteryx*. *Archaeopteryx* **15**, 107–108 (1997).
- Ostrom, J. H. The cursorial origin of avian flight. *Mem. Calif. Acad. Sci.* **8**, 73–81 (1986).
- Feduccia, A. Evidence from claw geometry indicating arboreal habits of *Archaeopteryx*. *Science* **259**, 790–793 (1993).
- Xu, X. & Wang, X.-L. New psittacosaur (Ornithischia, Ceratopsia) occurrence from the Yixian Formation of Liaoning, China and its stratigraphical significance. *Vertebrata Palasiatica* **36**, 147–158 (1998).
- Wellnhofer, P. Ein neues Exemplar von *Archaeopteryx*. *Archaeopteryx* **6**, 1–30 (1988).
- Xu, X., Wang, X.-L. & Wu, X.-C. A dromaeosaurid dinosaur with a filamentous integument from the Yixian Formation of China. *Nature* **401**, 262–266 (1999).

10. Russell, D. A. & Dong, Z.-M. A nearly complete skeleton of a new troodontid dinosaur from the Early Cretaceous of the Ordos Basin, Inner Mongolia, People's Republic of China. *Can. J. Earth Sci.* **30**, 2163–2173 (1994).
11. Currie, P. J. Bird-like characteristics of the jaws and teeth of troodontid theropods (Dinosauria, Saurischia). *J. Vert. Paleont.* **7**, 72–81 (1987).
12. Clark, J. M., Perle, M. & Norell, M. A. The skull of *Erlicosaurus andrewsi*, a Late Cretaceous "Segnosaur" (Theropoda: Therizinosauridae from Mongolia). *Am. Mus. Novit.* **3115**, 1–39 (1994).
13. Ostrom, J. H. Osteology of *Deinonychus antirrhopus*, an unusual theropod dinosaur from the Lower Cretaceous of Montana. *Peabody Mus. Nat. Hist. Bull.* **30**, 1–165 (1969).
14. Wellnhofer, P. Das siebte Exemplar von *Archaeopteryx* aus den Solnhofener Schichten. *Archaeopteryx* **11**, 1–47 (1993).
15. Currie, P. J. New information on the anatomy and relationships of *Dromaeosaurus albertensis* (Dinosauria: Theropoda). *J. Vert. Paleont.* **15**, 576–591 (1995).
16. Martin, L. D. in *Origins of the Higher Groups of Tetrapods* (eds Schultze, H.-P. & Trueb, L.) 485–540 (Comstock, Ithaca and London, 1991).
17. Zhou, Z. & Martin, L. D. Feathered dinosaur or bird?—a new look at the hand of *Archaeopteryx*. *Smithson. Contrib. Paleobiol.* **89**, 289–293 (1999).
18. Forster, C. A., Sampson, S. D., Chiappe, L. M. & Krause, D. W. The theropod ancestry of birds: new evidence from the Late Cretaceous of Madagascar. *Science* **279**, 1915–1919 (1998).
19. Makovicky, P. & Sues, H.-D. Anatomy and phylogenetic relationships of the theropod dinosaur *Microvenator celer* from the Lower Cretaceous of Montana. *Am. Mus. Novit.* **3240**, 1–27 (1998).
20. Holtz, T. R. Jr. The phylogenetic position of the Tyrannosauridae: implications for theropod systematics. *J. Paleont.* **68**, 1100–1117 (1994).
21. Norell, M. A. & Makovicky, P. J. Important features of the dromaeosaur skeleton: information from a new specimen. *Am. Mus. Novit.* **3215**, 1–28 (1997).
22. Currie, P. J. & Peng, J.-H. A juvenile specimen of *Saurornithoides mongoliensis* from the Upper Cretaceous of northern China. *Can. J. Earth Sci.* **30**, 2224–2230 (1994).
23. Currie, P. J., Rigby, J. K. & Sloan, R. E. in *Dinosaur Systematics: Perspectives and Approaches* (eds Carpenter, K. & Currie, P. J.) 108–125 (Cambridge Univ. Press, Cambridge, 1990).
24. Wellnhofer, P. A new specimen of *Archaeopteryx* from the Solnhofen Limestone. *Nat. Hist. Mus. Los Angeles County Sci. Ser.* **36**, 3–23 (1992).
25. Zhou, Z.-H. & Hou, L.-H. *Confuciusornis* and the early evolution of birds. *Vertebrata Palasiatica* **36**(2), 136–146 (1998).
26. Buffetaut, E., Suteethorn, V. & Tong, H.-Y. The earliest known tyrannosaur from the Lower Cretaceous of Thailand. *Nature* **381**, 689–691 (1996).
27. Xu, X., Tang, Z.-L. & Wang, X.-L. A therizinosaurid dinosaur with integumentary structures from China. *Nature* **399**, 350–354 (1999).
28. Clark, J. M., Hopson, J. A., Hernandez, R., Fastovsky, D. E. & Montellano, M. Foot posture in a primitive pterosaur. *Nature* **391**, 886–889 (1998).
29. Chiappe, L. M. Climbing *Archaeopteryx*? A response to Yalden. *Archaeopteryx* **15**, 109–112 (1997).
30. Chatterjee, S. *The Rise of Birds* (John Hopkins Univ. Press, Baltimore, 1997).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank D. Unwin, L. Chiappe, X.-C. Wu, J. Clark and L. Witmer for comments; M.-M. Chang and Y.-Q. Wang for assistance during the course of the work; H.-J. Wang for preparing the specimen; R.-S. Li for making the drawings; and J. Zhang and H.-L. You for taking the photographs. This work was supported by the Special Funds for Major State Basic Research Projects of China, and research grants from the National Geographic Society of the United States, Chinese Natural Science Foundation, and the Chinese Academy of Sciences.

Correspondence and requests for materials should be addressed to X.X. (e-mail: xxu@midwest.com.cn).

Mitochondrial genome variation and the origin of modern humans

Max Ingman*, Henrik Kaessmann†, Svante Pääbo† & Ulf Gyllenstein*

*Department of Genetics and Pathology, Section of Medical Genetics, Rudbeck Laboratory, University of Uppsala, S-751 85 Uppsala, Sweden

†Max Planck Institute for Evolutionary Anthropology, Inselstrasse 22, D-04103 Leipzig, Germany

The analysis of mitochondrial DNA (mtDNA) has been a potent tool in our understanding of human evolution, owing to characteristics such as high copy number, apparent lack of recombination¹, high substitution rate² and maternal mode of inheritance³. However, almost all studies of human evolution based on mtDNA sequencing have been confined to the control region, which constitutes less than 7% of the mitochondrial

genome. These studies are complicated by the extreme variation in substitution rate between sites, and the consequence of parallel mutations⁴ causing difficulties in the estimation of genetic distance and making phylogenetic inferences questionable⁵. Most comprehensive studies of the human mitochondrial molecule have been carried out through restriction-fragment length polymorphism analysis⁶, providing data that are ill suited to estimations of mutation rate and therefore the timing of evolutionary events. Here, to improve the information obtained from the mitochondrial molecule for studies of human evolution, we describe the global mtDNA diversity in humans based on analyses of the complete mtDNA sequence of 53 humans of diverse origins. Our mtDNA data, in comparison with those of a parallel study of the Xq13.3 region⁷ in the same individuals, provide a concurrent view on human evolution with respect to the age of modern humans.

The molecular clock hypothesis postulates that DNA sequence evolution is roughly constant over time in all evolutionary lineages. We used a test⁸ that compares the log likelihoods of trees reconstructed with and without the molecular clock assumption to examine the supposition that the mtDNA lineages evolve at 'clock-like' rates. The human mtDNA sequences, excluding the D-loop, have evolved at roughly constant rates ($P = 0.094$), and a relative rates test⁹, using a gorilla sequence as an outgroup, demonstrates that there is also no significant difference between the evolutionary rate of human and chimpanzee mtDNAs ($P = 0.123$), excluding the D-loop. In contrast, the D-loop has not evolved at a constant rate across all human lineages ($P < 0.001$), and is consequently less suitable for dating evolutionary events. Therefore, unless specifically mentioned, we have excluded the D-loop from the analyses that follow.

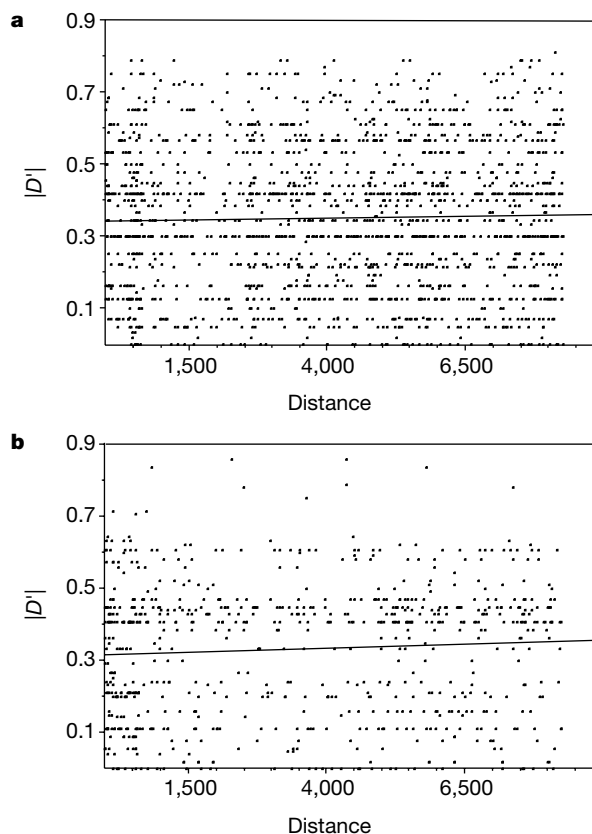


Figure 1 The relationship between linkage disequilibrium, measured by $|D'|$ versus distance between nucleotide sites for all 53 complete human mtDNA genomes. Values of ± 1.0 have been removed. **a**, Individuals of African descent ($n = 1,719$ comparisons), $R^2 = 0.001$; **b**, only non-African individuals ($n = 741$ comparisons), $R^2 = 0.005$.

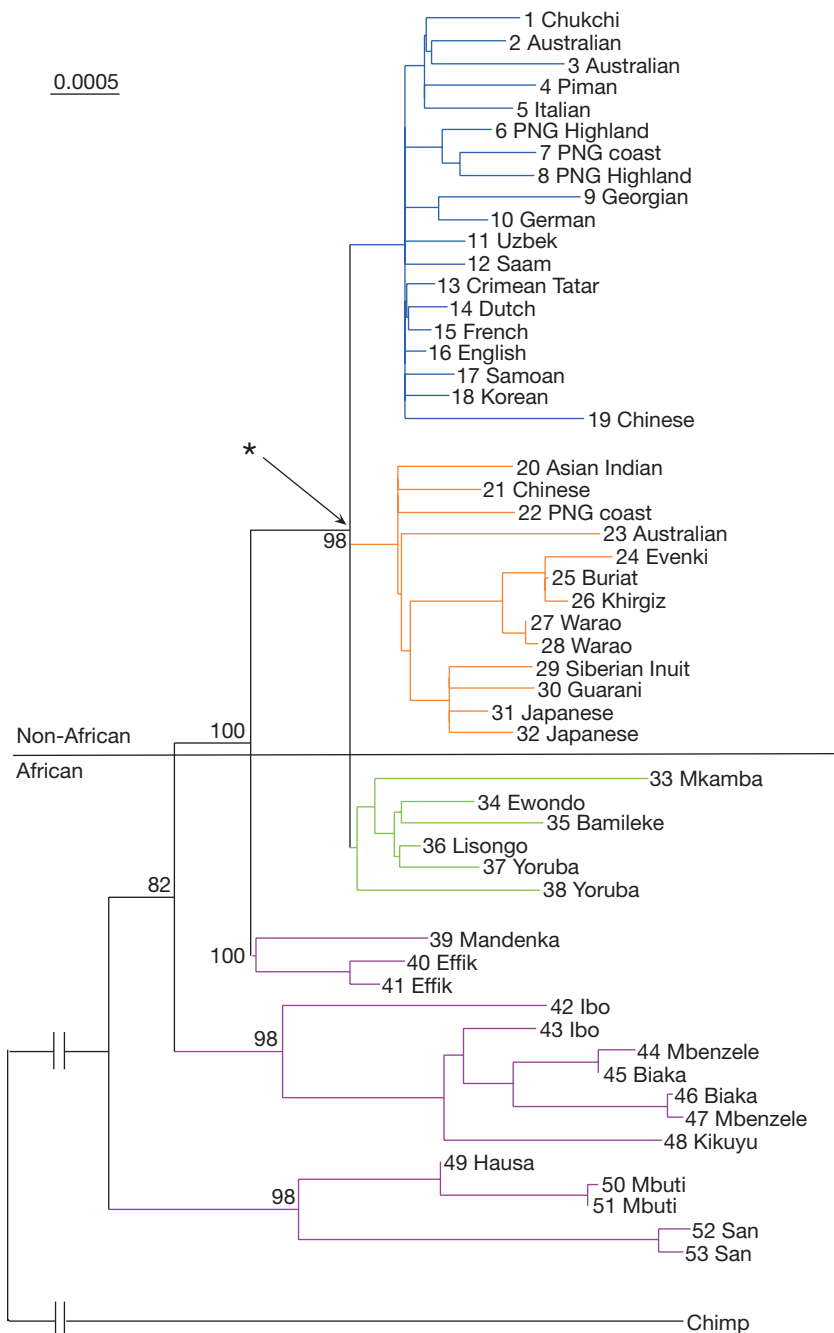


Figure 2 Neighbour-joining phylogram based on complete mtDNA genome sequences (excluding the D-loop). Data was constructed using PAUP*4.0 Beta (Sinauer Associates) and bootstrapped with 1,000 replicates (bootstrap values shown on nodes). The population origin of the individual is given at the twigs. Branches have been colour coded

as in Fig. 4. Individuals of African descent are found below the dashed line and non-Africans above. The node marked with an asterisk refers to the MRCA of the youngest clade containing both African and non-African individuals.

Table 1 Summary of statistical parameters for the mtDNA

	Data set	Length	<i>n</i>	<i>S</i>	MPSD	π
Total	All humans	16,553	53	657	61.1	3.7×10^{-3}
	Non-Africans	16,555	32	358	38.5	2.3×10^{-3}
	Africans	16,556	21	367	76.7	4.6×10^{-3}
D-Loop	All humans	1,118	53	141	17.2	1.5×10^{-2}
	Non-Africans	1,118	32	103	12.8	1.1×10^{-2}
	Africans	1,121	21	77	19.7	1.8×10^{-2}
Rest	All humans	15,435	53	516	43.9	2.8×10^{-3}
	Non-Africans	15,437	32	255	25.7	1.7×10^{-3}
	Africans	15,448	21	290	57.0	3.7×10^{-3}

mtDNA data are given as entire sequences, or separated into D-loop and rest (all but the D-loop), and further separated into groups of African and non-African individuals. Length, aligned sequence length excluding gaps; *n*, number of sequences; *S*, number of segregating sites; MPSD, mean pairwise sequence difference; and π , genetic diversity.

From the mean genetic distance between all the humans and the one chimpanzee sequence (0.17 substitutions per site) and the assumption, based on palaeontological¹⁰ and genetic¹¹ evidence, of a divergence time between humans and chimpanzees of 5 Myr, the mutation rate (μ) for the mitochondrial molecule, excluding the D-loop, is estimated to be 1.70×10^{-8} substitutions per site per year.

On the basis of the correlation between linkage disequilibrium and distance between sites, it has been claimed that mtDNA sequences show signs of recombination^{12,13}. These analyses can be criticized for the methodology used¹⁴, particularly for the use of a linkage disequilibrium measure (Hill and Robertson measure, r^2) that doesn't take allele frequency into account. We examined linkage

disequilibrium among all informative sites in our set of complete mtDNA genomes (including the D-loop) using a standard estimate (D') that allows all variable positions to be examined¹⁵. For this analysis, we studied the sequences of Africans and non-Africans separately. There is no correlation between D' and nucleotide distance between sites in the 53 sequences (African correlation coefficient, $R^2 = 1 \times 10^{-3}$; non-African $R^2 = 5 \times 10^{-3}$) (Fig. 1). We also analysed the association between r^2 and distance, and again there is no evidence of a correlation ($R^2 = 2.23 \times 10^{-6}$, and 1×10^{-3} , respectively; data not shown). Thus, it does not appear to be necessary to consider recombination as contributing to the evolution of mtDNA.

The two main hypotheses for the evolution of modern humans agree that *Homo erectus* spread from Africa around 2 Myr ago. The 'recent African origin' hypothesis^{16,17} states that anatomically modern humans originated in Africa 100,000–200,000 years ago and subsequently spread to the rest of the world, replacing archaic human forms with little or no genetic mixing. The alternative, 'multi-regional' hypothesis proposes that the transformation to anatomically modern humans occurred in different parts of the world, and supports this with fossil evidence of cultural and morphological continuity between archaic and modern humans outside Africa¹⁸. There are, of course, variants of these two basic hypotheses that introduce additional assumptions, such as gene flow, that make the hypotheses increasingly difficult to test.

Support for a recent African origin of modern humans has been provided by a number of mtDNA studies^{16,17,19,20}; however, these results have been troubled by the lack of statistical support for tree topology, especially the deep African branches^{21,22}. Lacking sufficiently strong empirical data, it is impossible to confidently place the root of modern human mtDNA lineages in sub-Saharan Africa. The neighbour-joining (NJ) tree constructed from our mtDNA sequences has a strongly supported basal branching pattern (Fig. 2). The three deepest branches lead exclusively to sub-Saharan mtDNAs, with the fourth branch containing both Africans and non-Africans. The deepest, statistically supported branch (NJ bootstrap = 100) provides compelling evidence of a human mtDNA origin in Africa.

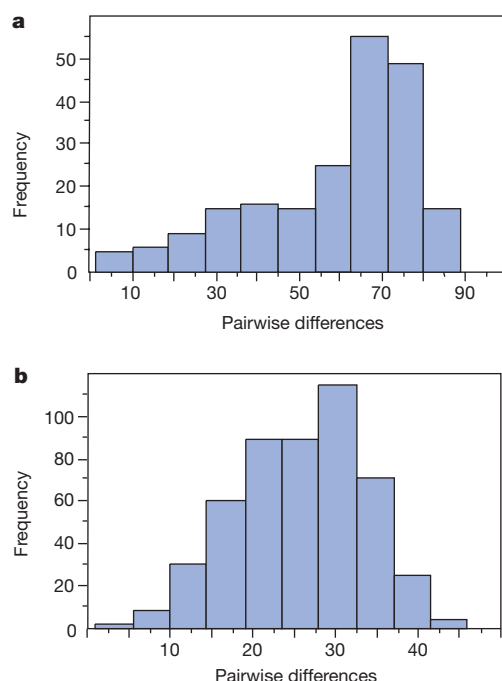


Figure 3 Mismatch distributions of pairwise nucleotide differences between mtDNA genomes (excluding the D-loop). **a**, African; **b**, non-African.

The amount of mtDNA sequence diversity (π) among Africans (3.7×10^{-3} nucleotide differences per site) is more than twice that among non-Africans (1.7×10^{-3}) (Table 1), corroborating earlier studies of the D-loop¹⁶ and nuclear loci²⁴. Also notable is the contrast between the deep branches of African mtDNAs and the 'star-like' phylogeny of non-African mtDNAs (Fig. 2). This high African diversity might result from either a considerably larger effective population size or a significantly longer genetic history. The 'star-like' phylogeny of the non-African sequences suggests a

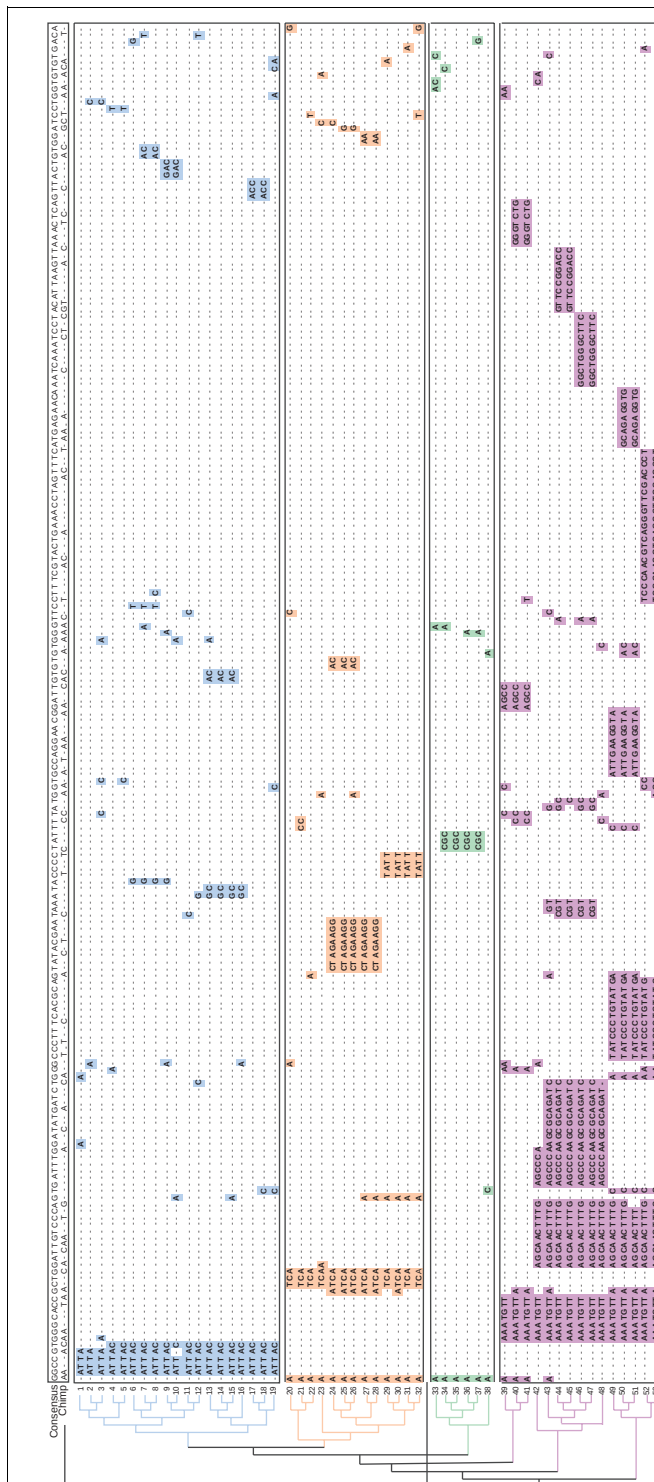
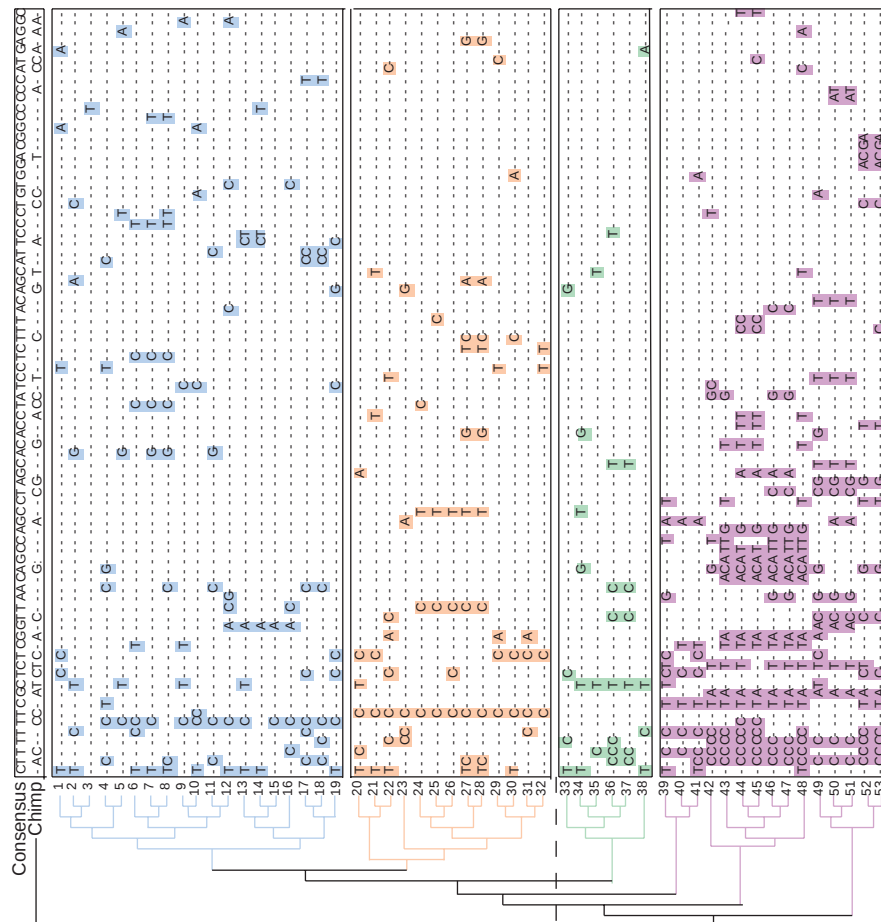


Figure 4 Data matrices showing all informative nucleotide positions, in decreasing order of frequency. Left, the whole mtDNA genome, excluding the D-loop. Right, the D-loop. The

population bottleneck, potentially associated with the colonization of Eurasia from Africa. The date of this exodus from Africa can be estimated if the departing group subsequently experienced a population expansion. The mtDNA mismatch distributions for Africans and non-Africans indicate a marked difference in population history for the two groups²⁵ (Fig. 3). Mitochondrial DNAs from individuals of African origin show a ragged distribution consistent with constant population size, whereas the bell-shaped distribution of the non-African comparisons clearly indicates a recent popula-

tion expansion. The assumption of constant population size can be verified by tests of selective neutrality that examine the correlation between the mean pairwise sequence difference (MPSD) and the number of segregating sites (S)^{26,27}. In the African group, we cannot reject this assumption (Fu and Li's $D = -1.17$ (ref. 26); Tajima's $D = -1.22$ (ref. 27)), consistent with the premise that the population has been of roughly constant size. However, it can be rejected in the non-African group ($D = -4.02$; $D = -2.28$), indicating that this group has experienced a period of population growth²⁵. The



trees on the left are cladograms with the same topology and numbering of individuals as the tree in Fig. 2. Individuals of African descent are found below the dashed line and

non-Africans above. The four major groups of sequences have been colour coded as in Fig. 2. Blocks denote groups of nucleotides that are identical in several sequences.

time when the expansion began was estimated ($\tau = 20.23$) to be about 1,925 generations ago²⁸. Assuming a generation time of 20 yr this equates to 38,500 yr BP, a date that coincides with the onset of a period of cultural change about 35,000–40,000 years ago²⁹. This involves, for example, the first appearance of regional cultural variation and the acceleration of artefactual change.

The age of the most recent common ancestor (MRCA) for mtDNA, on the basis of the maximum distance between two humans (5.82×10^{-3} substitutions per site between the Africans Mkamba and San), is estimated to be $171,500 \pm 50,000$ yr BP. We can also estimate the age of the MRCA for the youngest clade that contains both African and non-African sequences (Fig. 2, asterisk) from the mean distance of all members of that clade to their common node (8.85×10^{-4} substitutions per site) as $52,000 \pm 27,500$ yr BP. Because genetic divergence is expected to precede the divergence of populations, this date can be considered as the lower bound for an exodus from Africa.

Notably, a group of six African sequences (Fig. 4a, sequences 33–38) are genetically distant to those of other Africans, but share a common ancestor with non-Africans. These lineages represent descendants of a population that evidently gave rise to all the non-African lineages. Whether the ancestors of these six extant lineages originally came from a specific geographic region is not possible to determine, but we note that these sequences are from five populations that are now geographically unrelated.

Our study of the entire mitochondrial genome has some significant distinctions from previous studies of the D-loop. Most notably, the sequences outside of the D-loop evolve in a roughly 'clock-like' manner, enabling a more accurate measure of mutation rate, and therefore improved estimates of times to evolutionary events. Also of importance is the strong statistical support for the tree topology that has been lacking in earlier investigations. The difference between the D-loop and the rest of the molecule is visually evident in the contrast between the jumbled arrangement of polymorphic sites in the D-loop and the clear haplotypes defined by the sites in the rest of the molecule (Fig. 4).

The use of largely the same individuals in this study, as in that of the nuclear Xq13.3 (ref. 7) region, provides a unique opportunity to compare the information gained from the two genetic systems. Because the X chromosome has an effective population size three times that of mitochondria, the MRCA of an X-chromosomal locus is expected to be three times higher. Thus, the age of the MRCA of Xq13.3 is in agreement with the mtDNA data (mtDNA: 171,500; Xq13.3: 479,000 yr BP). The results from Xq13.3 also concur with the mtDNA data with respect to the greater genetic diversity found among African individuals relative to non-Africans, but Xq13.3 shows a considerably lower difference in diversity between the two groups (mtDNA: African, 3.9×10^{-3} , non-African, 1.7×10^{-3} ; Xq13.3: 3.5×10^{-4} , 3.05×10^{-4}). Owing to its lower substitution rate, only 33 segregating sites were present in 69 sequences of 10.2 kilobases (kb) at Xq13.3, as compared with the 657 segregating sites in the mitochondrial dataset. Comparisons of the mtDNA and Xq13.3 sequences carried by specific individuals show little correlation between the two loci, as expected from the different modes of inheritance. For example, the two Warao Indians showing the highest similarity in mtDNA have, in fact, two of the most divergent sequences studied at the Xq13.3 locus. Others, such as the Saami and Mandenka sequences, are closely related at Xq13.3 but have relatively high mitochondrial divergence.

Our results indicate that the field of mitochondrial population genomics will provide a rich source of genetic information for evolutionary studies. Nevertheless, mtDNA is only one locus and only reflects the genetic history of females. For a balanced view, a combination of genetic systems is required. With the human genome project reaching fruition, the ease by which such data may be generated will increase, providing us with an evermore detailed understanding of our genetic history. □

Methods

Sampling strategy and mtDNA sequences

To assess the global genetic diversity in humans, while analysing a restricted number of samples, we studied 53 individuals representing 14 of the major linguistic phyla. This sampling strategy attempts to avoid the bias inherent in selecting individuals on the basis of current world demographics, such as current population size or geographic location⁵. To provide an opportunity for comparison, we selected, where possible, the same individuals as those used by a previous study⁷ for the analysis of the Xq13.3 region. All the complete mtDNA sequences are unique and vary in length from 16,558 to 16,576 base pairs (bp). From nearly 900 kb sequenced, 5 heteroplasmic sites were confidently identified. We identified a total of 657 segregating sites (141 in the D-loop; 516 outside) among these 53 individuals, of which 283 (80 in the D-loop; 203 outside) showed the same polymorphism in at least 2 individuals (Fig. 4). The pairwise sequence distances between mtDNAs, corrected for multiple substitutions⁴, vary from 6.0×10^{-5} substitutions per site between two South American Indians (Warao) to 6.8×10^{-3} substitutions per site between two Africans (Mbenzele pygmy and San). The average distance between mtDNA genomes is 3.8×10^{-3} substitutions per site.

PCR primers and sequencing

The primers we used for polymerase chain reaction (PCR) amplification have been described³⁰. Sequencing was performed on the PCR products directly using BigDye (Applied Biosystems) chemistry. Separation of sequencing ladders was performed on the ABI 377 instrument for automated fragment analysis. We sequenced both forward and reverse strands. Sequence analysis was performed using Sequencing Analysis 3.3 (Applied Biosystems), and sequence alignment was made with Sequencher 3.1.1 (Gene Codes).

Accession numbers

The chimpanzee sequence (GenBank accession no. D38113 and the gorilla sequence (accession no. D38114) that we used as outgroups were obtained from a public mitochondrial sequence database (MITOMAP: <http://www.gen.emory.edu/mitomap.html>).

Received 9 June; accepted 30 August 2000.

1. Olivio, P. D., Van de Walle, M. J., Laipis, P. J. & Hauswirth, W. W. Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA D-loop. *Nature* **306**, 400–402 (1983).
2. Brown, W. M., George, M. Jr & Wilson, A. C. Rapid evolution of animal mitochondrial DNA. *Proc. Natl Acad. Sci. USA* **76**, 1967–1971 (1979).
3. Giles, R. E., Blanc, H., Cann, H. M. & Wallace, D. C. Maternal inheritance of human mitochondrial DNA. *Proc. Natl Acad. Sci. USA* **77**, 6715–6719 (1980).
4. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526 (1993).
5. Maddison, D. R., Ruvo, M. & Swofford, D. L. Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. *Syst. Biol.* **41**, 111–124 (1992).
6. Torroni, A. *et al.* mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am. J. Hum. Genet.* **62**, 1137–1152 (1998).
7. Kaessmann, H., Heissig, F., von Haeseler, A. & Paabo, S. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nature Genet.* **22**, 78–81 (1999).
8. Strimmer, K. & von Haeseler, A. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964–969 (1996).
9. Sarich, V. M. & Wilson, A. C. Generation time and genomic evolution in primates. *Science* **179**, 1144–1147 (1973).
10. Andrews, P. Evolution and environment in the Hominoidea. *Nature* **360**, 641–646 (1992).
11. Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).
12. Awadalla, P., Eyre-Walker, A. & Smith, J. M. Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science* **286**, 2524–2525 (1999).
13. Eyre-Walker, A., Smith, N. H. & Smith, J. M. How clonal are human mitochondria? *Proc. R. Soc. Lond. B* **266**, 477–483 (1999).
14. Kumar, S., Hedrick, P. & Stoneking, M. Questioning evidence for recombination in human mitochondrial DNA. *Science* **288**, 1931 (2000).
15. Lewontin, R. C. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**, 49–67 (1964).
16. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991).
17. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
18. Wolpoff, M. H. In *The Human Revolution: Behavioural and Biological Perspectives on the Origins of Modern Humans* (eds Mellars, P. & Stringer, C.) 62–108 (Princeton Univ. Press, Princeton, New Jersey, 1989).
19. Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. & Takahata, N. Recent African origin of modern humans revealed by complete sequences of hominid mitochondrial DNAs. *Proc. Natl Acad. Sci. USA* **92**, 532–536 (1995).
20. Ruvo, M. *et al.* Mitochondrial COII sequences and modern human origins. *Mol. Biol. Evol.* **10**, 1115–1135 (1993).
21. Templeton, A. R. Human origins and analysis of mitochondrial DNA sequences. *Science* **255**, 737 (1992).
22. Nei, M. Age of the common ancestor of human mitochondrial DNA. *Mol. Biol. Evol.* **9**, 1176–1178 (1992).
23. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

24. Zietkiewicz, E. *et al.* Genetic structure of the ancestral population of modern humans. *J. Mol. Evol.* **47**, 146–155 (1998).
25. Rogers, A. R. & Harpending, H. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9**, 552–569 (1992).
26. Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).
27. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
28. Rozas, J. & Rozas, R. DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. *Comput. Appl. Biosci.* **11**, 621–625 (1995).
29. Klein, R. G. *The Human Career: Human Biological and Cultural Origins* (Univ. Chicago Press, Chicago, 1989).
30. Reid, M. J., Taylor, S. L., Tobe, V. O. & Nickerson, D. A. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res.* **26**, 967–973 (1998).

Acknowledgements

We thank M. Stoneking for his advice regarding the analysis of recombination, and L. Cavalli-Sforza, G. Destro-Bisol, L. Excoffier, T. Jenkins, K. Kidd, J. Kidd, G. Klein, R. Mahabber, V. Nasidze, E. Poloni, H. Soodyall, M. Stoneking, M. Voevoda and S. Wells for samples. This work was supported by grants from Swedish Natural Sciences Research Council and Beijer Foundation.

Correspondence and requests for materials should be addressed to U.G. (e-mail: ulf.gyllensten@genpat.uu.se).

The *HIC* signalling pathway links CO₂ perception to stomatal development

Julie E. Gray*, Geoff H. Holroyd†, Frederique M. van der Lee‡, Ahmad R. Bahrami*, Peter C. Sijmons§, F. Ian Woodward||, Wolfgang Schuch¶ & Alistair M. Hetherington†

* Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN, UK

† Department of Biological Sciences, University of Lancaster, Lancaster LA1 4YQ, UK

‡ Zeneca Mogen, PO Box 628, 2300 AP Leiden, The Netherlands

§ Cellscreen, PO Box 134, 6700 AC Wageningen, Netherlands

|| Department of Animal and Plant Sciences, University of Sheffield, Sheffield, S10 2TN, UK

¶ Zeneca Wheat Improvement Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK

Stomatal pores on the leaf surface control both the uptake of CO₂ for photosynthesis and the loss of water during transpiration. Since the industrial revolution, decreases in stomatal numbers in parallel with increases in atmospheric CO₂ concentration have provided evidence of plant responses to changes in CO₂ levels caused by human activity^{1,2}. This inverse correlation between stomatal density and CO₂ concentration also holds for fossil material from the past 400 million years³ and has provided clues to the causes of global extinction events⁴. Here we report the

identification of the *Arabidopsis* gene *HIC* (for high carbon dioxide), which encodes a negative regulator of stomatal development that responds to CO₂ concentration. This gene encodes a putative 3-keto acyl coenzyme A synthase—an enzyme involved in the synthesis of very-long-chain fatty acids⁵. Mutant *hic* plants exhibit up to a 42% increase in stomatal density in response to a doubling of CO₂. Our results identify a gene involved in the signal transduction pathway responsible for controlling stomatal numbers at elevated CO₂.

As part of a promoter trap screen aimed at isolating tissue-specific genes in *Arabidopsis thaliana*^{6,7}, we identified an individual plant designated *hic* in which the β -glucuronidase (*GUS*) reporter gene was expressed only in the guard cells (Fig. 1) and nowhere else in the plant. Southern analysis of *hic* genomic DNA using a *GUS* gene probe detected a single band, suggesting that there was only one *GUS* gene insertion in this line (data not shown).

We next investigated the phenotype of the *hic* mutant and found it to be different from previously described *Arabidopsis* stomatal development mutants^{8,9}. *hic* stomata had no obvious phenotype and the plants were not wilting, suggesting that guard-cell function was not impaired. *hic* guard cells increased turgor in response to light and reduced turgor in response to the addition of the plant hormone abscisic acid in stomatal bioassays (data not shown). *hic* plants were indistinguishable from the parental ecotype when analysed by infrared thermography (data not shown); however, analysis of plants that had been grown under differing CO₂ concentrations showed that the *HIC* gene affects stomatal development.

We grew *hic* and C24, the parental ecotype, at elevated and ambient concentrations of CO₂, and measured the stomatal index and density on the leaf surface. A reduction in stomatal index and density in response to elevated CO₂ is well characterized in some species, including three ecotypes of *Arabidopsis*^{1,2}, and indicates that the plant has the potential for increased water-use efficiency with increasing CO₂. Table 1 shows that growth of *hic* plants under elevated CO₂ induced marked increases in both stomatal index and

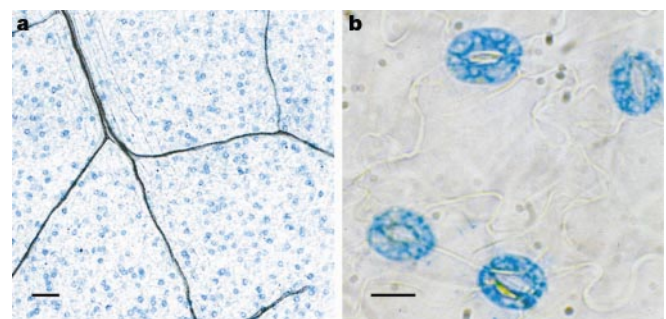


Figure 1 *GUS* expression in *hic* *Arabidopsis* plants is specific to guard cells. Part of a *hic* leaf histochemically stained for *GUS* activity shows guard-cell-specific expression at low (a; scale bar, 100 μ m) and higher magnification (b; scale bar, 10 μ m).

Table 1 Effect of ambient and elevated CO₂ on stomatal and epidermal cell density and stomatal index

Experiment	Line	Stomatal density (n mm ⁻²)				Epidermal cell density (n mm ⁻²)				Stomatal index (%)			
		Ambient	Elevated	Difference (<i>P</i>)	% change	Ambient	Elevated	Difference (<i>P</i>)	% change	Ambient	Elevated	Difference (<i>P</i>)	% change
1	C24	280 (9)	233 (8)	<0.001	-17.0	1,047 (9)	902 (8)	<0.001	-13.8	21.1 (9)	20.5 (8)	NS	-3.1
	<i>hic</i>	231 (8)	320 (8)	<0.001	+38.0	891 (8)	902 (8)	NS	+1.2	20.8 (8)	26.0 (8)	<0.005	+25
2	C24	265 (7)	245 (9)	NS	-7.8	1,056 (7)	1,028 (9)	NS	-2.7	20.0 (7)	19.1 (9)	NS	-4.4
	<i>hic</i>	257 (9)	324 (10)	<0.001	+25.2	1,047 (9)	1,061 (10)	NS	+1.3	19.8 (9)	23.2 (10)	<0.001	+17.8
3	C24	324 (9)	245 (7)	<0.001	-24.5	1,322 (9)	1,058 (7)	<0.005	-20	19.6 (9)	18.7 (7)	NS	-4.7
	<i>hic</i>	273 (7)	387 (9)	<0.001	+41.8	1,053 (7)	1,092 (9)	NS	+3.7	20.5 (7)	26.2 (9)	<0.001	+27.6
4	C24	273 (12)	283 (12)	NS	+3.1	1,052 (12)	1,049 (12)	NS	-0.3	20.6 (12)	21.2 (12)	NS	+3.2
	<i>hic</i>	266 (12)	351 (12)	<0.001	+31.8	1,035 (12)	1,062 (12)	NS	+2.6	20.6 (12)	24.9 (12)	<0.001	+21.2

Data are presented for four separate experiments in the *hic* mutant and the parental ecotype (C24). The average density or stomatal index is presented for each treatment. The numbers in brackets indicate the number of individual plants in each treatment. Treatments were compared using the Mann-Whitney Rank Sum Test (SigmaStat 2.03). NS, not significant.

density. By contrast, CO₂ enrichment induced no significant changes in stomatal index and either reductions or no significant change in stomatal density of the C24 control plants. Although the C24 ecotype of *Arabidopsis* appears relatively insensitive to CO₂, other ecotypes of this species show reductions in stomatal density in response to elevated CO₂ (ref. 2). These data suggested that the *HIC* gene product is involved in a signal transduction pathway in which elevated CO₂ influences stomatal development.

To confirm that the disruption of *HIC* was responsible for the increase in stomatal index observed at elevated CO₂, we adopted two independent experimental procedures. First, we showed that the *GUS* insertion in *hic* plants co-segregated with the increased stomatal index phenotype when these plants were grown at elevated CO₂. T2 generation *hic* seeds (from a seed population that segregated 3:1 on selective media) were germinated without selection and the plants were grown under elevated CO₂. A leaf that had developed during growth at elevated CO₂ was removed from each plant and scored for stomatal index and the presence of the *GUS* gene was determined for each plant using polymerase chain reaction (PCR) and a histochemical *GUS* assay. All plants containing the *GUS* transgene had *GUS* activity and had significantly higher stomatal indices than those that lacked this insert (Fig. 2). *HIC* does not act as a recessive gene, and these results are consistent with a dominant-negative or gene-dosage effect.

Second, we tried to replicate the *hic* phenotype by manipulating the expression of the *HIC* gene. We transformed the C24 ecotype with a gene construct designed to express antisense *HIC* mRNA (see below for complementary DNA identification). When three independently transformed lines of T3 generation antisense plants were

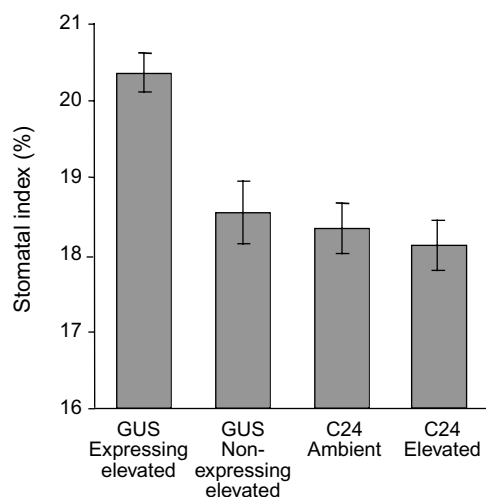


Figure 2 The *GUS* gene co-segregates with the elevated CO₂ stomatal index phenotype. T2 *hic* seeds were germinated on non-selective media before transfer to the elevated CO₂ chamber for 22 days. Of the 20 plants, 16 showed *GUS* expression and 4 did not, on the basis of a histochemical *GUS* assay^{22,23}. Bars represent standard error. $P < 0.001$ between *GUS* expressers and *GUS* non-expressers. A similar set of data was obtained in a second, separate run of this experiment (not shown).

grown under elevated CO₂, they all showed significant increases in stomatal index (Table 2). PCR with reverse transcription of RNA (RT-PCR) indicated that these antisense lines had reduced levels of *HIC* messenger RNA (see Methods). The results from this experiment show that when the expression of *HIC* is reduced plants have increased stomatal index and density at elevated CO₂. Both these results provide strong support for our hypothesis that the increase in stomatal index observed when *hic* plants are grown at elevated CO₂ is caused by a disruption in *HIC*.

To study the nature of the *HIC* gene, we isolated a region of DNA from the *hic* genome adjacent to the inserted *GUS* gene. This *hic* DNA fragment was used to identify genomic and cDNA *HIC* gene fragments from non-tagged *Arabidopsis*. Analysis of the nucleotide sequences showed that the *GUS* gene had inserted in *hic* plants slightly downstream from a predicted open reading frame in the 3' untranslated region of a transcribed gene. It was also apparent that the *HIC* gene was disrupted by a fragment of mitochondrial DNA introducing three stop codons. This would be expected to result in the production of a severely truncated gene product (Fig. 3). On the basis of nucleotide and derived peptide sequence homologies, *HIC* was found to be most similar to the *Arabidopsis KCS1* gene that encodes a 3-ketoacyl coenzyme A (CoA) synthase¹⁰ (KCS; Fig. 3). As the insertion of the fragment of mitochondrial DNA into the *HIC* gene is expected to terminate translation in the predicted active site of the KCS enzyme¹¹ (Fig. 3), it seems likely that the activity of the *HIC*-encoded putative KCS will be considerably reduced.

Our data indicate that the activity of a guard-cell KCS, the condensing enzyme in the microsomal fatty-acid elongase complex¹⁰, may be involved in the stomatal patterning response to elevated CO₂. Fatty-acid elongases catalyse the synthesis of very long chain fatty acids, and in plants they are involved in the synthesis of waxes, glycerolipids, sphingolipids and cutin⁵. In *hic* plants, it is possible that the lesion in this putative KCS prevents the synthesis of component(s) of the extracellular matrix found at the guard-cell surface. Our hypothesis to account for the *hic* phenotype is that loss of the component(s) synthesized by *HIC* results in a modification of stomatal development in response to CO₂.

Other, independent evidence lends support to our hypothesis. We reasoned that as *HIC* encodes a putative KCS and KCS enzymes are involved in wax biosynthesis, plants carrying lesions in wax biosynthesis genes should display aberrant stomatal densities when compared with wild-type plants. In fact, almost 30 years ago, the barley wax-deficient *eceriferum-g* mutant was reported to display abnormal stomatal patterning¹². We examined stomatal indices in the *cer1* and *cer6* wax-accumulation mutants of *Arabidopsis*^{13,14}. With respect to their parental ecotype, both *cer1* and *cer6* displayed greatly increased stomatal indices (Table 3). These results show that alterations in leaf wax biosynthesis in *Arabidopsis* are associated with aberrant stomatal densities, in support of our hypothesis. In the case of *hic*, however, the disruption to wax biosynthesis must be subtle, as a scanning electron microscopic comparison of the epidermis of *hic* and C24 revealed no obvious differences in wax morphology (data not shown).

The link between the putative guard-cell KCS gene *HIC* and alterations in stomatal density is supported by work on another

Table 2 Effect of ambient and elevated CO₂ on stomatal and epidermal cell densities and stomatal indices

Line	Stomatal density (n mm ⁻²)				Epidermal cell density (n mm ⁻²)				Stomatal index (%)			
	Ambient	Elevated	Difference (P)	% change	Ambient	Elevated	Difference (P)	% change	Ambient	Elevated	Difference (P)	% change
C24	242 (9)	252 (11)	NS	+4.1	990 (9)	1,088 (11)	NS	+9.9	19.6 (9)	18.8 (11)	NS	-4.2
Antisense 1	235 (9)	286 (10)	<0.001	+21.7	1,028 (9)	1,062 (10)	NS	+3.3	18.6 (9)	21.2 (10)	<0.001	+13.9
Antisense 2	240 (10)	272 (10)	<0.001	+13.3	1,010 (10)	997 (10)	NS	-1.3	19.2 (10)	21.5 (10)	<0.001	+11.9
Antisense 3	243 (10)	263 (9)	<0.03	+8.2	929 (10)	907 (9)	NS	-2.4	20.7 (10)	22.5 (9)	<0.001	+8.7

Comparisons between *Arabidopsis* plants transformed with the *HIC* gene in an antisense orientation and the C24 ecotype were done by the Mann-Whitney Rank Sum Test (SigmaStat 2.03). NS, not significant.

Arabidopsis KCS gene called *FIDDLEHEAD* (*FDH*) (Fig. 3). Mutations in *FDH* manifest themselves in a range of developmental responses^{15,16}, however, from the perspective of our results, the most interesting observation is that mutations in *FDH* alter leaf epidermal cell fate and specifically control trichome development¹⁷. Given that trichomes and guard cells both develop from proto-dermal cells¹⁸, this is very strong evidence that the product(s) of KCS enzymes are capable of altering the fate of cells in the leaf epidermis.

Our results indicate that *hic* plants carry a mutation in a KCS gene that results in a disruption in the signal transduction pathway responsible for the control of stomatal patterning in response to elevated CO₂. The simplest explanation, consistent with existing data on KCS¹⁶, to account for the *hic* phenotype is that disruption of the putative KCS encoded by *HIC* results in altered permeability of the guard-cell extracellular matrix. These alterations in guard-cell extracellular matrix permeability then alter the diffusion of an elevated CO₂-stimulated morphogen responsible for the control of stomatal development. As the normal response to elevated CO₂ is either to reduce stomatal index and density or to keep them unchanged (Table 1), it seems likely that this morphogen is a negative regulator of guard-cell development. These results are in line with the lateral inhibition hypothesis for the control of stomatal development^{19,20} and are consistent with current ideas on the role of cell-cell communication in stomatal patterning¹⁸. To our knowledge, *HIC* is the first gene to be identified that affects the developmental responses of plants to global changes in atmospheric composition. □

Methods

Plant material and growth conditions

We used the following plant material: *hic*, *cer1*, *cer6* (refs 13, 14) and ecotypes. Vernalized seeds were germinated on 0.5× Murashige and Skoog basal salts, 1% (w/v) sucrose, 0.6% (w/v) agar at 22–24 °C under a 10 h/14 h light (150 μmol m⁻² s⁻¹)/ dark regime. Where selection was required, media included 10 mg l⁻¹ hygromycin. To investigate the effects of elevated CO₂, resistant plants were selected from a non-homozygous T2 *hic* line (segregating roughly 3:1 resistant:susceptible on hygromycin) and transferred to a peat-based compost mix at 16–18 days old. C24 and *hic* plants were then grown in a matched pair of chambers built using the hot gas bypass system as described²¹ for 22–27 days at 22 °C ± 0.5 day, 22 °C ± 0.5 night, 10 h/14 h light/dark regime, 60% relative humidity, 180 μmol m⁻² s⁻¹ lighting. CO₂ concentration was controlled (elevated 1,000 p.p.m. ± 100 and ambient 350–480 p.p.m.) and monitored by an infrared gas analyser (WMA-2, PP Systems) switched between chambers by a solenoid valve on a 30-s cycle.

GUS staining

Leaves were histochemically stained for the presence of the GUS gene product as described²² followed by a clearing step²³.

Determination of stomatal indices

A leaf that had developed during the course of the experiment was removed from each plant and stomatal density, epidermal density and stomatal index determined as described²⁴ from the abaxial surface of three areas of each leaf using the dental rubber impression technique^{25,26}. Leaves were selected on the basis of uniform comparable size between treatments and across experiments.

Analysis of the *HIC* gene sequence

hic DNA 5' to the *GUS* gene insertion was isolated by inverse PCR essentially as described²⁷ using *pfu* polymerase, oligonucleotide primers 5'-CAGAACTTACGTACACTTTTC-3' and 5'-CATCTTCTTCTATGCCCTACTC-3', and *EcoRV*-restricted *hic* genomic DNA, and ligated into pGEMT (Promega) to produce pFL44. This was used as a probe to screen a

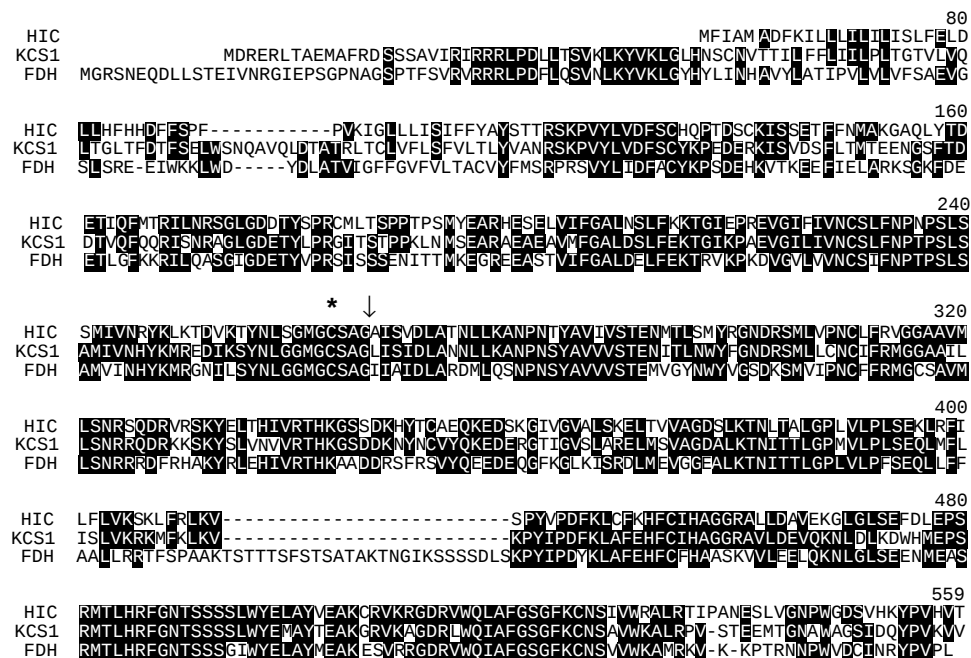


Figure 3 *HIC* is similar to *Arabidopsis* 3-keto acyl CoA synthases. Alignment of the predicted peptide sequence of *HIC* with those of KCS¹⁰ (accession number AF053345) and FDH^{15–17} (accession number AJ010713) with identical residues highlighted in black

and the tentatively identified active-site cysteine¹¹ marked with an asterisk. The arrow indicates the position of the mitochondrial DNA insertion that disrupts the *HIC* gene in the region of the putative active site.

Table 3 Stomatal indices of *cer1* and *cer6* *eceriferum* mutants of *Arabidopsis* and their parental ecotype

Genetic background	Stomatal index (%)	Difference (<i>P</i>)	Relative % change
Ler	11.8 (8)	—	—
<i>cer1</i>	16.9 (7)	<0.005	+43.2
<i>cer6</i>	15.4 (15)	<0.005	+30.5

The numbers in brackets indicate the number of plants examined. Comparison with parental ecotype, landsberg *erecta* (Ler) was made using Student's *t*-test.

C24 *Arabidopsis* genomic DNA library and isolate a *Bgl*II genomic DNA fragment encoding the complete *HIC* open reading frame, and also to identify ATTS5501 expressed sequence tag (clone YAY1019) by a BLASTN homology search of the dbest database (this cDNA was provided by J. Giraudat, CNRS, Gif, France). The complete nucleotide sequences of ATTS5501 (2,073 bp), pFL30 (5,600 bp) and pFL44 (1,584 bp) inserts were determined. The *HIC* gene is identical to a region of BAC T3A4 (accession no. AC005819) on *Arabidopsis* chromosome 2. The cloned sequences pFL30 and ATTS5501 are identical in their regions of overlap. pFL44 contains an additional 123-bp mitochondrial DNA insertion from nucleotide position 482–604 (identical to part of the *Arabidopsis* mitochondrial genome sequence part B (accession number Y08502) which is also present elsewhere on chromosome 2; ref. 28). It is probable that the insertion causing the *hic* mutation occurred during the *Agrobacterium-tumefaciens*-mediated transformation process used to generate *hic*. PCR was used to confirm the presence of this mitochondrial DNA insertion in *hic* plants and its absence in C24. The *GUS* gene is inserted 89-bp 3' of the putative *HIC* stop codon. RT–PCR was used to confirm the presence of the *GUS* insert downstream of *hic* in the same gene transcript by amplifying a fragment of DNA that spanned the two coding regions.

Generation of *HIC* antisense plants.

ATTS5501 cDNA was excised from pBluescript using *Eco*RI and ligated into pART7 (ref. 29). The resulting plasmid was digested with *Nco*I to release the cDNA in the reverse orientation between the CaMV 35S RNA promoter and OCS 3' terminator which was ligated into pART27 (ref. 29). Gene constructs were confirmed by DNA sequencing and *Agrobacterium*-mediated transformation was carried out as described³⁰. The presence of the transgene in plants was verified by PCR (data not shown). RT–PCR with primers (5'-GCTAGTGGTGAACGTCATGC-3' and 5'-ACAAAATCGTTACGCAAG-3' designed specifically to amplify a 1,281-bp region of the 5' untranslated portion of the *HIC* mRNA that differs from other known *KCS*-like genes) was used to show that the level of the *HIC* gene transcript was either considerably reduced in the antisense plants (line AS3) or undetectable by this method (AS1 and AS2), in comparison with C24 plants, which gave a clear band of DNA of the expected size on agarose gel electrophoresis. Separate amplification reactions with ubiquitin-specific primers were carried out to confirm equal amounts of mRNA. Control reactions minus reverse transcriptase gave no signal.

Received 15 May; accepted 23 October 2000.

- Woodward, F. I. Stomatal numbers are sensitive to CO₂ increases from pre-industrial levels. *Nature* **327**, 617–618 (1987).
- Woodward, F. I. & Kelly, C. K. The influence of CO₂ concentration on stomatal density. *New Phytol.* **131**, 311–327 (1995).
- McElwain, J. C. & Chaloner, W. G. Stomatal density and index of fossil plants track atmospheric carbon-dioxide in the paleozoic. *Ann. Bot.* **76**, 389–395 (1995).
- McElwain, J. C., Beerling, D. J. & Woodward, F. I. Fossil plants and global warming at the Triassic–Jurassic boundary. *Science* **285**, 1386–1390 (1999).
- Post-Beittenmiller, D. Biochemistry and molecular biology of wax production in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47**, 405–430 (1996).
- Topping, J. F., Wei, W. B. & Lindsey, K. Functional tagging of regulatory elements in the plant genome. *Development* **112**, 1009–1019 (1991).
- Goddijn, O. J. M., Lindsey, K., van der Lee, F. M., Klap, J. C. & Sijmons, P. C. Differential gene-expression in nematode-induced feeding structures of transgenic plants harbouring GUSA fusion constructs. *Plant J.* **4**, 863–873 (1993).
- Yang, M. & Sack, F. D. The *too many mouths* and *four lips* mutations affect stomatal production in *Arabidopsis*. *Plant Cell* **7**, 2227–2239 (1995).
- Berger, D. & Altman, T. A subtilisin-like serine protease involved in the regulation of stomatal density and distribution in *Arabidopsis thaliana*. *Genes Dev.* **14**, 1119–1131 (2000).
- Todd, J., Post-Beittenmiller, D. & Jaworski, J. G. *KCS1* encodes a fatty acid elongase 3-ketoacyl-CoA synthase affecting wax biosynthesis in *Arabidopsis thaliana*. *Plant J.* **17**, 119–130 (1999).
- Lassner, M. W., Lardizabal, K. & Metz, J. G. A *jojoba* β-keto-CoA synthase cDNA complements the canola fatty acid elongation mutation in transgenic plants. *Plant Cell* **8**, 281–292 (1996).
- Zeiger, E. & Stebbins, L. Developmental genetics in barley: a mutant for stomatal development. *Am. J. Bot.* **59**, 143–148 (1972).
- Jenks, M. A., Tuttle, H. A., Eigenbrode, S. D. & Feldmann, K. A. Leaf epicuticular waxes of the *eceriferum* mutants in *Arabidopsis*. *Plant Physiol.* **108**, 369–377, (1995).
- Post-Beittenmiller, D. The cloned *Eceriferum* genes of *Arabidopsis* and the corresponding *Glossy* genes in maize. *Plant Physiol. Bioch.* **36**, 157–166 (1998).
- Lolle, S. J., Cheung, A. Y. & Sussex, I. M. *fiddlehead*: an *Arabidopsis* mutant constitutively expressing an organ fusion program that involves interactions between epidermal cells. *Dev. Biol.* **152**, 383–392 (1992).
- Lolle, S. J. et al. Developmental regulation of cell interactions in the *Arabidopsis fiddlehead-1* mutant: a role for the epidermal cell wall and cuticle. *Dev. Biol.* **189**, 311–321 (1997).
- Yephremov, A. et al. Characterization of the *FIDDLEHEAD* gene of *Arabidopsis* reveals a link between adhesion response and cell differentiation in the epidermis. *Plant Cell* **11**, 2187–2201 (1999).
- Larkin, J. C., Marks, M. D., Nadeau, J. & Sack, F. Epidermal cell fate and patterning in leaves. *Plant Cell* **9**, 1109–1120 (1997).
- Bünning, E. & Sagermeyer, H. Die Bildung des Spaltöffnungsmusters in der Blattepidermis. *Z. Naturforsch.* **3b**, 203–216 (1948).
- Korn, R. W. Evidence in dicots for stomatal patterning by inhibition. *Int. J. Plant Sci.* **154**, 367–377 (1993).
- Neighbour, E. A. et al. A small-scale controlled environment chamber for the investigation of effects of pollutant gases on plants growing at cool or sub-zero temperature. *Environ. Pollution* **64**, 155–168 (1990).
- Jefferson, R. A. Assaying chimeric genes in plants: The *GUS* gene fusion system. *Plant Mol. Biol. Rep.* **5**, 387–405 (1987).
- Beckman, A. A. & Engler, A. A. An easy technique for the clearing of histochemically stained plant tissue. *Plant Mol. Biol. Rep.* **12**, 37–42 (1994).

- Salisbury, E. J. On the causes and ecological significance of stomatal frequency with special reference to woodland flora. *Phil. Trans. R. Soc. Lond. B* **216**, 1–65 (1927).
- Weyers, J. D. B. & Johansen, L. G. Accurate estimation of stomatal aperture from silicone rubber impressions. *New Phytol.* **101**, 109–115 (1985).
- Poole, L., Weyers, J. D. B., Lawson, T. & Raven, J. A. Variations in stomatal density and index: implications for paleoclimatic reconstructions. *Plant Cell Environ.* **19**, 705–712 (1996).
- Barthels, N. et al. Regulatory sequences of *Arabidopsis* drive reporter gene expression on nematode feeding structures. *Plant Cell* **9**, 2119–2134 (1997).
- Lin, X. Y. et al. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402**, 761–765 (1999).
- Gleave, A. P. A versatile binary vector system with a T-DNA organisational structure conducive to efficient integration of cloned DNA into the plant genome. *Plant Mol. Biol.* **20**, 1203–1207 (1992).
- Valvekens, D., Van Montagu, M. & Van Lijsebettens, M. *Agrobacterium tumefaciens*-mediated transformation of *Arabidopsis thaliana* root explants by using kanamycin selection. *Proc. Natl Acad. Sci. USA* **85**, 5536–5540 (1988).

Acknowledgements

We acknowledge the assistance of J. Proctor, J. Balk, A. G. Moir and P. Tripathi with this work. We would like to thank K. Lindsey and S. Bright for helpful discussions.

Correspondence and requests for materials should be addressed to A.M.H. (e-mail: A.Hetherington@lancaster.ac.uk). The accession numbers for ATTS 5501, FL30 and FL44 are in GenBank under accession numbers AF188484, AF188485 and AF188486, respectively.

The *ELF3* *zeitnehmer* regulates light signalling to the circadian clock

Harriet G. McWatters*, Ruth M. Bastow*, Anthony Hall & Andrew J. Millar

Department of Biological Sciences, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK

* These authors contributed equally to this work

The circadian system regulates 24-hour biological rhythms¹ and seasonal rhythms, such as flowering². Long-day flowering plants like *Arabidopsis thaliana*, measure day length with a rhythm that is not reset at lights-off³, whereas short-day plants measure night length on the basis of circadian rhythm of light sensitivity that is set from dusk². *early flowering 3* (*elf3*) mutants of *Arabidopsis* are photoperiodic⁴ and exhibit light-conditional arrhythmia^{5,6}. Here we show that the *elf3-7* mutant retains oscillator function in the light but blunts circadian gating of *CAB* gene activation, indicating that deregulated phototransduction may mask rhythmicity. Furthermore, *elf3* mutations confer the resetting pattern of short-day photoperiodism, indicating that gating of phototransduction may control resetting. Temperature entrainment can bypass the requirement for normal *ELF3* function for the oscillator and partially restore rhythmic *CAB* expression. Therefore, *ELF3* specifically affects light input to the oscillator, similar to its function in gating *CAB* activation, allowing oscillator progression past a light-sensitive phase in the subjective evening. *ELF3* provides experimental demonstration of the *zeitnehmer* ('time-taker') concept^{7,8}.

As *elf3* mutants are rhythmic in darkness (DD)^{5,6}, *ELF3* cannot be an essential component of the circadian oscillator. We tested whether the apparent arrhythmia in constant light (LL) of *elf3* plants was in fact masking an underlying oscillation. We entrained wild-type, null mutant *elf3-1* and partial mutant *elf3-7* (refs 5, 6) plants to light/dark cycles (LD), transferred them to LL and released replicate samples into DD at 2-h intervals, monitoring the phase of *CAB* expression in DD to determine the state of the oscillator in the preceding LL interval. We reasoned that if an underlying oscillator in the *elf3* mutants was masked by constant illumination, its phase should be reflected in the phase of the peak in DD. If the oscillator were dysfunctional in *elf3* mutants under LL, the timing of the *CAB*

peak would be determined by the time at which the plants were transferred to darkness.

In wild-type plants, the timing of the *CAB* peak after the light to dark transition remains under circadian control (Fig. 1a), occurring at or immediately preceding the phases predicted from the entraining LD cycle (roughly 4, 28 or 52 h after lights-on). The circadian system was stably entrained to the LD cycles, continued to oscillate during the 2 d of LL and was little affected by the last light to dark transition. *elf3-1* shows no evidence of circadian regulation after 10 h or more of light (Fig. 1a), as the phase of peak *CAB* expression is determined by the time of transfer to DD. The peak occurs after 9–12 h in darkness, indicating that the oscillator resumes from a phase equivalent to 8–11 h after lights-on. This pattern is strikingly similar to the photoperiodic response rhythm of short-day plants². *elf3-7* showed an intermediate phenotype, retaining circadian-regulated *CAB* expression with a phase closer to wild type for up to 14 h of LL (Fig. 1b; up to 20 h in a minority of plants, data not shown). After this interval, the transfer to DD determined the peak phase, as in *elf3-1*. Partial *ELF3* function (in *elf3-7*) therefore maintained rhythmicity during the first 14 h of LL. The absence of *ELF3* (in *elf3-1*) resulted in a circadian oscillator that functioned still more briefly and abnormally in LL. We revealed a functional oscillator in *elf3-7* during the first 14 h of LL, by testing circadian rhythms in DD, whereas directly testing *CAB* expression under LL showed complete arrhythmia⁶. We conclude that the *CAB* output rhythm is masked during the first subjective day in LL, but that the oscillator is dysfunctional thereafter.

Continuous acute activation of *CAB* expression in response to continuous light might cause such masking. We therefore directly monitored the acute response to light in *elf3* mutants grown in LD 12/12 and transferred to DD (Fig. 2). Replicate samples that were

exposed to 20 min light at 2-h intervals showed a rhythm of responsiveness to light in the wild type, with little or no acute response during the first and second subjective nights and large responses during the subjective day (as previously reported⁹). The *elf3* mutants had lost the normal circadian gating, showing strong acute responses throughout the first subjective night. In *elf3-1* in particular, the level of the acute response is increased during the subjective day. A low-amplitude modulation of the acute response is preserved during the first subjective night, but after 30 h the *CAB* activation levels are constant in both *elf3* mutants, indicating a complete absence of circadian gating. The gating defect of *elf3-7* suggests that the misregulated acute response activated *CAB* continuously in LL, masking the functional circadian oscillator. The extent of masking of the circadian system might be determined through the expression of genes such as *CCR2*, which are regulated by the clock but not light.

The high-amplitude rhythms of *CAB* expression in wild-type plants under LL therefore depend on circadian gating of light responsiveness in the late subjective day and early subjective night. The term *zeitnehmer* ('time-taker') has been applied to an input pathway rhythmically regulated by feedback from an oscillator, thus creating rhythmic input even under constant conditions^{10,11}. The gating defect of even the partial loss-of-function *elf3-7* mutant suggests that high levels of *ELF3* are required for a normal gating function. The *elf3* mutations also result in a non-sustainable oscillator in LL, with a partial defect allowing progression to the mid-subjective night and the severe loss-of-function of *elf3-1* losing rhythmic control more rapidly. We propose that *ELF3* does not function directly in the circadian oscillator, but gates light input to the oscillator, allowing oscillator progression through a light-sensitive phase around dusk. A defect in the circadian gating of

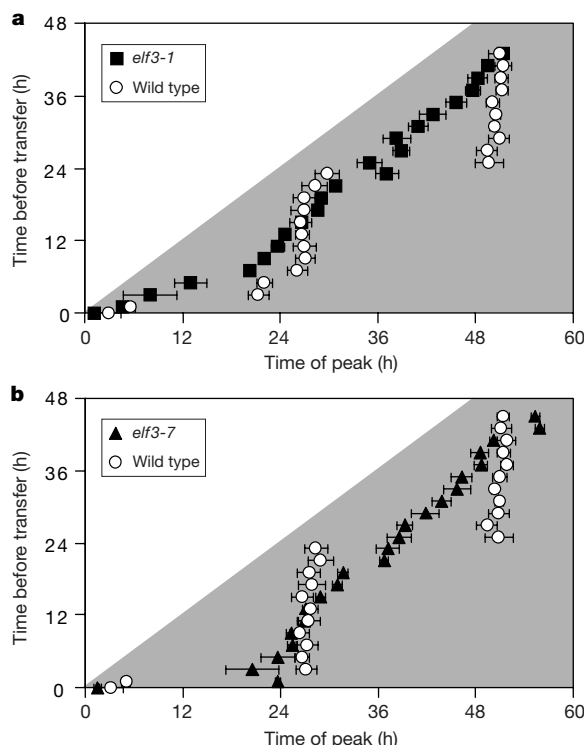


Figure 1 Timing of *CAB*–*LUC* peaks after LD cycles. **a**, *elf3-1* and its wild-type parent, gl1. **b**, *elf3-7* and its wild-type parent, 424. All plants were entrained to LD 12/12 for 6 d. At ZT 0, subjective dawn, plants were transferred to bright LL at 22 °C. Plants were transferred to darkness at 2-h intervals starting at subjective dawn; scintillation counter assays began at the time of transfer. White area represents the length of time plants spent in constant light before being transferred to the dark. Data are representative of two independent experiments; mean $\pm$ s.e.m. is shown, $n = 10$ –16.

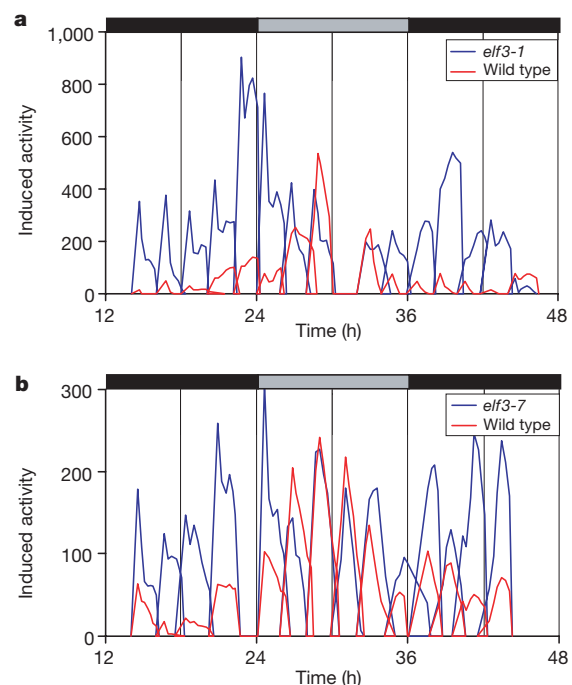


Figure 2 Circadian modulation of the acute response to light in *elf3* mutants. **a**, *elf3-1* and its wild-type parent, gl1. **b**, *elf3-7* and its wild-type parent, 424. All plants were entrained to LD 12/12 for 6 d. At ZT 12 on day 6, plants were transferred to DD. Replicate samples were then exposed to a 20-min white-light pulse every 2 h. Luminescence was measured both before and after the light treatment. The mean resulting induction of *CAB*–*LUC* luminescence was calculated by subtraction of the basal luminescence level of individual seedlings ($n = 24$) before the light treatment. Light and dark shading represents subjective day and subjective night, respectively. Data are representative of three independent experiments.

phototransduction can account for both the masking and oscillator arrest phenotypes of *elf3* mutants. As *elf3* mutations are sufficient to confer strong resetting at lights off, the *ELF3 zeitnehmer* might in part determine the mode of daylength measurement. Our evidence for an oscillation in *elf3* plants under LL, albeit transitory and at an early phase, suggests that these mutants might respond differentially to short and very short photoperiods. We suggest that the *elf3* oscillator arrests in LL rather than oscillating in an altered state and being reset by the light to dark transition; the pattern of phases in DD does not show the periodic variation indicative of a light limit cycle².

Reasoning that a masked oscillator might be revealed by reducing the strength of the masking light signal, we entrained plants using LD cycles of low fluence light ($10 \mu\text{mol m}^{-2} \text{s}^{-1}$) and assayed *CAB-LUC* activity in this fluence rate LL. Only wild-type plants exhibited free-running circadian rhythms: reducing fluence rate alone was insufficient to restore overt rhythmicity to *elf3* mutants (see Supplementary Information). We next replaced light with temperature entrainment as temperature cycles, but not light cycles, can entrain *frq* null strains of *Neurospora crassa*⁸. In low fluence light, wild-type plants entrained to temperature cycles. When released into constant conditions, plants continued rhythmically in a free-run from the original phase, with lower mean levels and lengthened period ($\tau = 25\text{--}26 \text{ h}$) (Fig. 3a), as described for dim LL¹². Although circadian rhythms in both *elf3-1* and *elf3-7* plants could be driven by temperature cycles in LL they did not subsequently produce free-running circadian rhythms in bright or dim LL (Fig. 3a; and Supplementary Information); neither temperature nor light alone could rescue circadian behaviour under constant

illumination.

As temperature cycles did not restore overt circadian rhythms of *CAB* expression in *elf3* under LL, we tested the phase of rhythms in DD to reveal the behaviour of the underlying oscillator. We entrained plants in temperature cycles in bright LL ($140 \mu\text{mol m}^{-2} \text{s}^{-1}$), then released them into DD at constant temperature. As after light entrainment, wild-type plants showed circadian *CAB* expression peaks phased with respect to the (discontinued) temperature cycle, even after two days in constant conditions (Fig. 4). Thus, the wild-type oscillator was entrained by temperature and was not reset by a single light–dark transition. After transfer to darkness at the first subjective dawn, the phase of both *elf3-1* and *elf3-7* was identical to that of wild-type plants (Fig. 4). Previous experiments with LD cycles at constant temperature have shown that *elf3-7* has an earlier peak than wild type, with that of *elf3-1* being earlier still^{5,6}. Thus temperature cycles, in contrast to light cycles (Fig. 2; and see ref. 6), entrained the *elf3* plants initially to a wild-type phase relationship.

This intriguing result implies that the *elf3* lesion specifically affects the light input pathway. During the first 5 h in constant conditions *elf3-1* showed an early phase when transferred to darkness, with a sharp jump to a later phase in plants transferred after 7–11 h, implying that an oscillator is still working for the first subjective day in constant conditions. Thereafter, *elf3-1* plants did not show circadian regulation: peaks of *CAB-LUC* expression were phased from the transfer to darkness (Fig. 4a), as after LD cycles (Fig. 2), suggesting the oscillator was no longer functional. In contrast to *elf3-1* and unlike its response to LD cycles, *elf3-7* showed a stable circadian pattern up to 48 h in LL (Fig. 4b). Thus an oscillator remains in *elf3-7* which may be entrained by tempera-

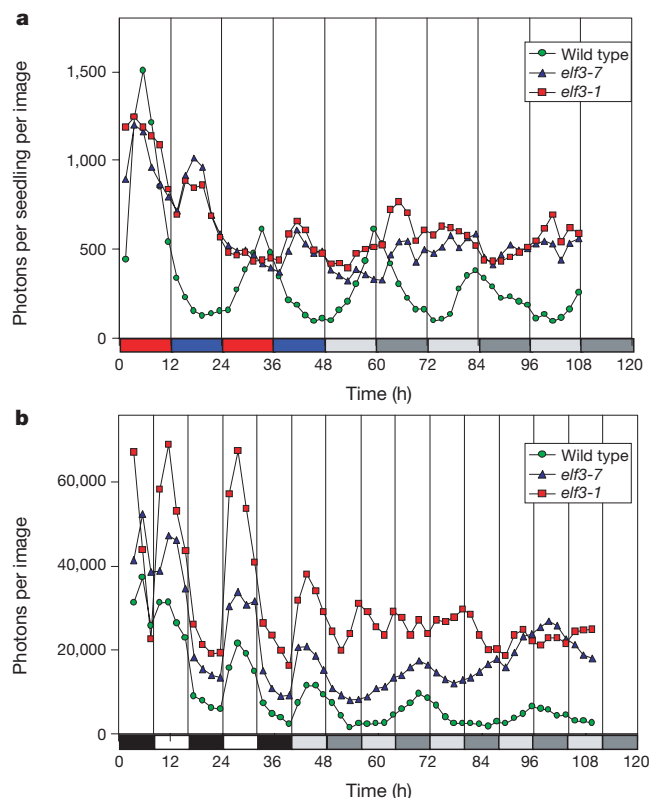


Figure 3 Response to temperature or combined light and temperature cycles. **a**, *CAB-LUC* activity in temperature cycles in low fluence LL ($10 \mu\text{mol m}^{-2} \text{s}^{-1}$). T-cycle = 24 h: 'day' is 12 h at 24 °C, 'night' is 12 h at 18 °C. After measuring luminescence for 48 h, temperature cycles were discontinued and plants were placed in dim LL, 22 °C for the remainder of the assay. Data are representative of four independent experiments. Red and blue bars represent warm and cold periods, respectively; light and dark shading represents subjective day and subjective night, respectively. **b**, *CAB-LUC* activity in

simultaneous light and temperature cycles. T-cycle = 16 h: 'day' is 8 h dim light ($10 \mu\text{mol m}^{-2} \text{s}^{-1}$) at 24 °C, 'night' is 8 h dark at 18 °C. After T-cycle entrainment, plants were transferred to dim LL, 22 °C at subjective dawn. Data are representative of two independent experiments. White and black bars represent light and darkness, respectively. Light and dark shading represents subjective day and subjective night, respectively.

ture cycles and free-run in constant conditions (Fig. 3a). The output of this covert oscillator is easily masked, even by dim light, leading to overt arrhythmia. In contrast, the *elf3-1* oscillator stops within 12 h of constant light.

We reasoned that since circadian timing in *elf3-7* is restored after temperature cycles, light and temperature cycles together could reinforce circadian function. We combined these two *zeitgebers* to produce warm days and cooler nights. We used various entraining periods (T-cycles) as the earlier phase of *elf3* after LD cycles can be interpreted as a short endogenous period. No genotype free-ran after 12-h T-cycles (data not shown). Wild-type plants showed free-running rhythmicity after 16-h T-cycles (Fig. 3b) or after 24-h T-cycles (see Supplementary Information). Entrainment to combined low light and temperature cycles of 16, 20 or 24 h restored free-running rhythmicity in a minority of samples of *elf3-1* and *elf3-7* plants (Fig. 3b). In two independent experiments, 3 out of 28 samples of *elf3-1* and 10 out of 29 samples of *elf3-7* free-ran after such cycles with periods of around 26 h (judged free-running if FFT-NLLS analysis returned a period of 15–35 h with a relative amplitude error less than 0.6; ref. 13) (Fig. 3b). In comparison, after entrainment to a single *zeitgeber*, either light or temperature, none out of 40 *elf3-1* samples and 1 out of 40 *elf3-7* samples free-ran (Fig. 3a; and Supplementary Information). In contrast, all wild-type controls entrained to a single *zeitgeber* were rhythmic (40 out of 40 samples); 47 out of 57 samples were rhythmic after combined light and temperature entrainment. The free-running periods obtained in *elf3* were not distinguishable from those of wild-type plants: the long period is attributed to the low light intensity. Reinforcing the entraining effects of temperature with low-amplitude light cycles can overcome the masking effects of light on the circadian system in *elf3* plants to reveal the underlying oscillator despite ongoing illumination.

Our results indicate that despite their common arrhythmia in LL^{5,6}, the circadian system of *elf3-7* is less compromised than that of

elf3-1. We suggest *elf3-1* is arrhythmic because its circadian system arrests rapidly in LL, although it is capable of restarting from a constant phase in darkness. In contrast, *elf3-7* plants retain sufficient ELF3 function to maintain (masked) oscillations for at least 2 d in constant light. Their oscillator can be rescued and free-running rhythmicity restored by prior entrainment by temperature, especially if masking is reduced by low light intensity. The different responses of *elf3-7* to light or temperature cycles indicate that the *elf3* lesion affects a light input pathway, rather than the temperature-sensitive oscillator. The gating of the acute response of *CAB* in the early to mid night (*zeitgeber* time (ZT) 12–16) coincides with the arrest of the oscillator in *elf3-7* (circa ZT 14) after light and *elf3-1* (circa ZT 11) after temperature cycles; this is also the phase from which their oscillators restart in darkness. This strongly implies that *ELF3* functions at this phase of the circadian cycle. A rhythmic *ELF3*-dependent antagonism of the phototransduction pathway to the clock and to *CAB* at this phase can account for all the *elf3* phenotypes. Our data provide empirical evidence for the *zeitnehmer* ('time-taker') concept¹⁰: we show that *ELF3* is a part of the *zeitnehmer* feedback loop necessary to maintain self-sustained oscillations rather than an oscillator component. □

Methods

Plant materials and growth conditions

Transgenic *A. thaliana* containing the *CAB-LUC* fusion together with the *elf3-1* mutation in the Columbia *gl1* background or with the *elf3-7* mutation in the Columbia-0 background have been described^{5,6}. In all cases wild type refers to the respective parent. Seeds were imbibed then stratified (48 h at 4 °C) on solid Murashige and Skoog (MS) medium with 3% sucrose¹⁴.

Entrainment conditions

After imbibition and stratification as described above¹⁴, seedlings were entrained for 6 d in light/dark and/or temperature cycles as appropriate in controlled environment chambers (Percival or Sanyo) before measurements began. Constant light or LD cycles involved high fluence fluorescent light (140 $\mu\text{mol m}^{-2} \text{s}^{-1}$) (Figs 1 and 4) or low fluence fluorescent light (10 $\mu\text{mol m}^{-2} \text{s}^{-1}$) (Fig. 3). Light pulses were 80 $\mu\text{mol m}^{-2} \text{s}^{-1}$ (Fig. 2). Constant temperature was 22 °C. Temperature cycles consisted of warm periods (24 °C) and cold periods (18 °C).

Bioluminescence assay

Luminescence levels of individual seedlings were measured in a Packard Topcount¹⁵. Luminescence of groups of seedlings ($n = 30$ –100) was measured by low-light video imaging using a liquid nitrogen cooled camera (Princeton Instruments)^{16,17} and normalized for seedling number⁹. Instrument background has been subtracted from the data presented.

Received 30 June; accepted 1 September 2000.

- Dunlap, J. C. Molecular bases for circadian clocks. *Cell* **96**, 271–290 (1999).
- Lumsden, P. J. In *Biological Rhythms and Photoperiodism in Plants* (eds Lumsden, P. J. & Millar, A. J.) 167–181 (BIOS Scientific, Oxford, 1998).
- Samach, A. & Coupland, G. Time measurement and the control of flowering in plants. *BioEssays* **22**, 38–47 (2000).
- Zagotta, M. T., Shannon, S., Jacobs, C. & Meeks-Wagner, D. R. Early-flowering mutants of *Arabidopsis thaliana*. *Aust. J. Plant Physiol.* **19**, 411–418 (1992).
- Hicks, K. A. et al. Conditional circadian dysfunction of the *Arabidopsis* early-flowering 3 mutant. *Science* **274**, 790–792 (1996).
- Reed, J. W. et al. Independent action of ELF3 and phyB to control hypocotyl elongation and flowering time. *Plant Physiol.* **122**, 1149–1160 (2000).
- Roenneberg, T. & Merrow, M. Molecular circadian oscillators: An alternative hypothesis. *J. Biol. Rhythms* **13**, 167–179 (1998).
- Merrow, M., Brunner, M. & Roenneberg, T. Assignment of circadian function for the *Neurospora* clock gene frequency. *Nature* **399**, 584–586 (1999).
- Millar, A. J. & Kay, S. A. Integration of circadian and phototransduction pathways in the network controlling *CAB* gene transcription in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **93**, 15491–15496 (1996).
- Roenneberg, T. & Merrow, M. Circadian systems and metabolism. *J. Biol. Rhythms* **14**, 449–459 (1999).
- Lakin-Thomas, P. L. Circadian rhythms: new functions for old clock genes? *Trends Genet.* **16**, 135–142 (2000).
- Somers, D. E. & Quail, P. H. Phytochrome-mediated light regulation of *PHYA*- and *PHYB*-GUS₂ transgenes in *Arabidopsis thaliana* seedlings. *Plant Physiol.* **107**, 523–534 (1995).
- Dowson-Day, M. J. & Millar, A. J. Circadian dysfunction causes aberrant hypocotyl elongation patterns in *Arabidopsis*. *Plant J.* **17**, 63–71 (1999).
- Millar, A. J., Carré, I. A., Strayer, C. A., Chua, N. H. & Kay, S. A. Circadian clock mutants in *Arabidopsis* identified by luciferase imaging. *Science* **267**, 1161–1163 (1995).
- Carré, I. A. & Kay, S. A. Multiple DNA–protein complexes at a circadian-regulated promoter element. *Plant Cell* **7**, 2039–2051 (1995).

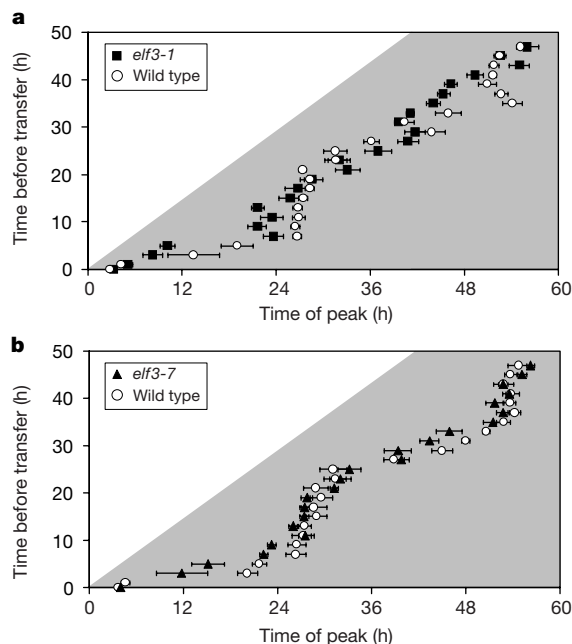


Figure 4 Timing of *CAB-LUC* peaks after temperature cycles. **a**, *elf3-1* and its wild-type parent, *gl1*. **b**, *elf3-7* and its wild-type parent, 424. All plants were entrained to 12 h at 24 °C, 12 h of 18 °C temperature cycles in high fluence LL (140 $\mu\text{mol m}^{-2} \text{s}^{-1}$) for 6 d. From ZT 0, subjective dawn, plants were maintained in bright LL at 22 °C. Plants were transferred to darkness at 2-h intervals starting at subjective dawn; scintillation counter assays began at the time of transfer. White area represents the length of time plants spent in constant light before being transferred to the dark. Data are representative of two independent experiments; mean $\pm$ s.e.m. is shown, $n = 10$ –16.

16. Bogner, L. K. *et al.* The circadian clock controls the expression pattern of the circadian input photoreceptor, phytochrome B. *Proc. Natl. Acad. Sci. USA* **96**, 14652–14657 (1999).
17. Millar, A. J., Short, S. R., Chua, N. H. & Kay, S. A. A novel circadian phenotype based on firefly luciferase expression in transgenic plants. *Plant Cell* **4**, 1075–1087 (1992).

Supplementary information is available on Nature's World-Wide web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

R.M.B. was supported by a Gatsby graduate studentship; this work was supported by grants from the BBSRC to A.J.M.

Correspondence and requests for materials should be addressed to A.J.M. (e-mail: Andrew.Millar@warwick.ac.uk).

μ -Opioid receptor desensitization by β -arrestin-2 determines morphine tolerance but not dependence

Laura M. Bohn*, Raul R. Gainetdinov*, Fang-Tsyr Lin†, Robert J. Lefkowitz† & Marc G. Caron*

* Howard Hughes Medical Institute Laboratories, Departments of Cell Biology and Medicine, Duke University Medical Center, Durham, North Carolina 27710, USA

† Howard Hughes Medical Institute Laboratories, Departments of Biochemistry and Medicine, Duke University Medical Center, Durham, North Carolina 27710, USA

Morphine is a powerful pain reliever, but also a potent inducer of tolerance and dependence. The development of opiate tolerance occurs on continued use of the drug such that the amount of drug required to elicit pain relief must be increased to compensate for diminished responsiveness^{1–3}. In many systems, decreased responsiveness to agonists has been correlated with the desensitization of G-protein-coupled receptors. *In vitro* evidence indicates that this process involves phosphorylation of G-protein-coupled receptors and subsequent binding of regulatory proteins called β -arrestins^{4,5}. Using a knockout mouse lacking β -arrestin-2 (β arr2^{-/-}), we have assessed the contribution of desensitization of the μ -opioid receptor to the development of morphine antinociceptive tolerance and the subsequent onset of physical dependence. Here we show that in mice lacking β -arrestin-2, desensitization of the μ -opioid receptor does not occur after chronic morphine treatment, and that these animals fail to develop antinociceptive tolerance. However, the deletion of β -arrestin-2 does not prevent the chronic morphine-induced up-regulation of adenylyl cyclase activity, a cellular marker of dependence, and the mutant mice still become physically dependent on the drug.

We have shown previously⁶ that morphine-induced antinociception is enhanced and prolonged in mice lacking β -arrestin-2. This perpetuation of morphine analgesia suggests that mice lacking β -arrestin-2 may be resistant to the desensitization of the morphine signal. As specific inhibitors of desensitization do not exist, this genetically altered animal model provides a means to potentially abrogate desensitization of the μ -opioid receptor (μ OR) in response to morphine. Thus, we examined the regulation of the μ OR in relationship to the development of morphine tolerance and dependence in β arr2^{-/-} mice.

An acute challenge with a high dose of morphine induces acute morphine antinociceptive tolerance 24 h after the initial challenge. In this scheme, mice are assessed for their nociceptive response

latencies after a moderate dose of morphine (10 mg per kg, subcutaneously (s.c.)) 24 h after receiving an injection of either saline or 100 mg per kg morphine. Indicative of antinociceptive tolerance, wild-type mice exhibited a roughly 50% reduction in morphine responsiveness if they had received morphine, as compared to saline, the day before (Fig. 1a). The β arr2^{-/-} mice maintained the same degree of responsiveness to morphine, however, whether they had been treated with saline or morphine on the previous day. It therefore seemed that the β arr2^{-/-} mice did not develop acute antinociceptive tolerance to morphine.

As, in the clinical setting, tolerance to morphine's analgesic properties usually develops over continued use of moderate levels of the drug, we evaluated the analgesia provided after daily administration of morphine. Mice were injected daily with morphine, and paw-withdrawal latencies were recorded (Fig. 1b). Although the wild-type littermates had significantly diminished responsiveness to the drug by day 5, the knockout mice continued to experience as much antinociception on day 5 to day 9 as on day 1. To characterize this resistance to morphine antinociceptive tolerance in the β arr2^{-/-}

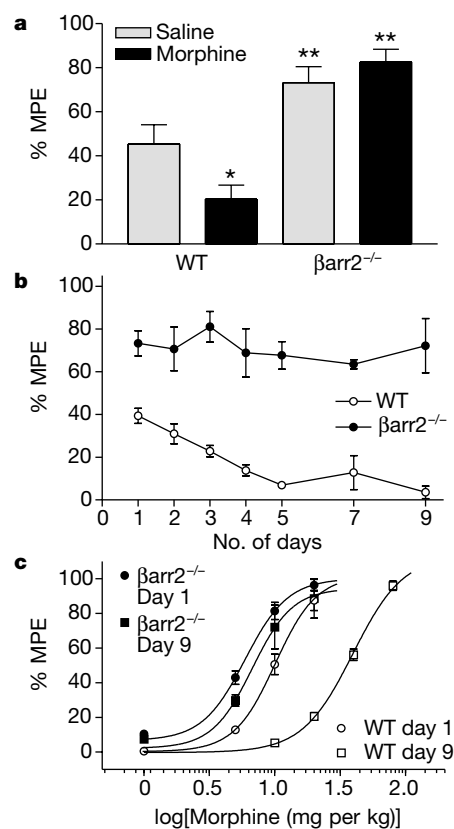


Figure 1 Lack of morphine antinociceptive tolerance in β arr2^{-/-} mice. **a**, Acute tolerance. Wild-type (WT) and β arr2^{-/-} mice were treated with either saline or morphine (100 mg per kg, s.c.). After 24 h, mice were treated with morphine (10 mg per kg, s.c.) and hot-plate-response latencies (56 °C) were recorded. Data are presented as the mean $\pm$ s.e.m., $n = 7-8$. Asterisk, $P < 0.05$ versus WT + saline; two asterisks, $P < 0.05$ versus WT, Student's *t*-test. **b**, Chronic tolerance. Mice were treated daily with morphine (10 mg per kg, s.c.); antinociception was assessed 30 min after the injection on the days indicated. Means $\pm$ s.e.m. are shown. $P < 0.0001$, WT versus β arr2^{-/-}; two-way analysis of variance (ANOVA). **c**, Chronic tolerance. Dose-response curves were determined using a cumulative dosing scheme on day 1 and day 9. On day 1, both genotypes were treated with 1, 5, 10 and 20 mg per kg, s.c. On day 9, β arr2^{-/-} mice ($n = 6-12$) were again challenged with this same dose scheme, whereas WT ($n = 7-14$) were cumulatively treated with 10, 20, 40 and 80 mg per kg, s.c. and β arr2^{-/-}. Means $\pm$ s.e.m. are shown. ED₅₀ (50% effective dose) values were calculated by nonlinear regression analysis (GraphPad Prism); 95% confidence intervals: day 1: WT, 10.1 (8.4–12.1); β arr2^{-/-}, 5.9 (5.0–7.0); day 9: WT, 39.6 (34.0–46.1); β arr2^{-/-}, 6.7 (4.8–9.3).

mice, we examined the effectiveness of morphine in a dose-response scheme, comparing responsiveness on day 1 to that on day 9. Notably, the β arr2^{-/-} mice did not experience a shift in their sensitivity to morphine after chronic daily administration, whereas the wild-type mice experienced a significant rightward shift in efficacy after continued treatment (Fig. 1c).

Both the maintenance of morphine responsiveness in the β arr2^{-/-} mice during chronic administration and the proposed role of β -arrestin-2 in desensitization of G-protein-coupled receptors suggest that a loss of receptor desensitization may underlie the resistance to the development of tolerance. As the analgesic actions of morphine are mediated predominantly through the μ OR⁷⁻⁹, we focused on the desensitization state of this receptor in the β arr2^{-/-} mice. To evaluate

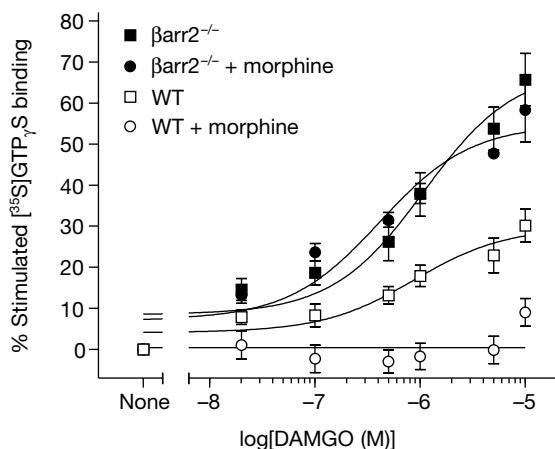


Figure 2 [³⁵S]GTPγS binding to brainstem membranes from β arr2^{-/-} and WT mice before and after chronic morphine treatment. After 5 d of chronic morphine (10 mg per kg, s.c. per day) brain regions were dissected from WT and β arr2^{-/-} mice. Data were analysed by nonlinear regression (GraphPad Prism) and are presented as the mean $\pm$ s.e.m. of at least three experiments performed in triplicate in which WT and β arr2^{-/-} brain regions were assayed simultaneously. In the absence of agonist stimulation, basal [³⁵S]GTPγS binding was WT, 2,049 $\pm$ 91 c.p.m.; β arr2^{-/-}, 1,964 $\pm$ 113 c.p.m. The curves were compared by one-way ANOVA followed by a Bonferroni post-hoc test (GraphPad Prism). β arr2^{-/-} and β arr2^{-/-} + morphine curves do not differ but are significantly greater than WT and WT + morphine curves ($P < 0.001$). WT + morphine differs from WT ($P < 0.001$).

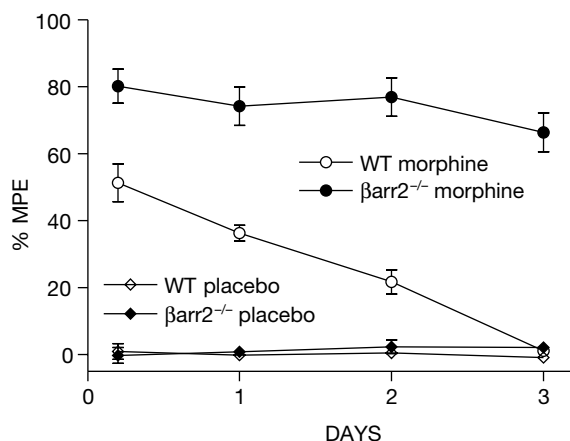


Figure 3 Lack of antinociceptive tolerance in β arr2^{-/-} mice after 72 h of morphine-pellet implantation. Morphine or placebo pellets were implanted subcutaneously and the degree of antinociception was determined by measuring latency of hot-plate (56 °C) responses as in Fig. 1. WT ($n = 8-12$) and knockout ($n = 7-9$) animals were analysed in parallel during the same experiment. Means $\pm$ s.e.m. are shown. WT + morphine versus β arr2^{-/-} + morphine, $P < 0.0001$, two-way ANOVA.

the contribution of this state of the μ OR to the development of behavioural tolerance, we examined the ability of the receptor to couple to G protein on stimulation with the μ OR-selective agonist [*D*-Ala₂, (MePhe)₄, Gly₅-ol]-enkephalin (DAMGO) after chronic administration of morphine. In mice treated only with saline, there is already a difference in the extent of DAMGO-stimulated G-protein coupling in brainstem membranes between the two genotypes, as indicated by [³⁵S]GTPγS binding (Fig. 2). The elevated [³⁵S]GTPγS binding in the β arr2^{-/-} mice is similar to that reported in periaqueductal grey membranes from naive wild-type and β arr2^{-/-} mice⁶. After 5 d of chronic morphine treatment, when wild-type mice experienced substantial nociceptive tolerance, mice were killed and brain regions were used for biochemical studies. The coupling of the μ OR with G proteins was maintained in brainstem membranes from β arr2^{-/-} mice after chronic morphine treatment, whereas the receptors were uncoupled in the membranes from the chronically treated wild-type mice (Fig. 2). Similar results were also obtained in periaqueductal grey membranes (data not shown). The number of μ ORs, as assessed by [³H]naloxone binding⁶ in brainstem-membrane preparations from these same tissues did not decrease after treatment, nor did they vary between genotypes (wild type + saline, 49 $\pm$ 7; wild type + morphine, 57 $\pm$ 3; β arr2^{-/-} + saline, 54 $\pm$ 9; β arr2^{-/-} + morphine, 53 $\pm$ 1 fmol per mg protein). Similar results were also obtained in a separate set of brainstem-membrane-binding experiments using the μ OR-selective agonist, [³H]-DAMGO (data not shown).

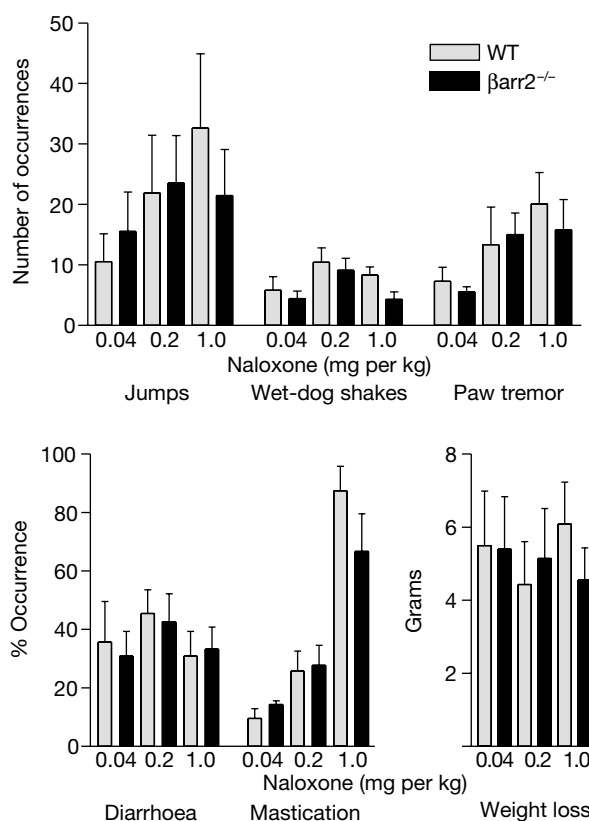


Figure 4 Naloxone-precipitated withdrawal after 72 h of morphine pellet implantation. Naloxone (0.04, 0.2 or 1.0 mg per kg, i.p.) was administered, and the number of jumps, wet-dog shakes and paw-tremor events were counted over 30 min. Occurrence of diarrhoea and mastication was noted as present or absent in six 5-min intervals and normalized according to a maximum of six possible (0 = 0%, 6 = 100%). The weight of each animal was determined before and 30 min after the naloxone injection. There were no significant differences ($P < 0.05$, Student's *t*-test) between the two genotypes. Data are presented as the mean $\pm$ s.e.m.; WT, $n = 7-11$; β arr2^{-/-}, $n = 7-9$. Different sets of animals were used for each dose of naloxone.

To study further this loss of the development of morphine tolerance in the $\beta\text{arr}2^{-/-}$ mice, we adopted a model of chronic morphine treatment in which the mice were continuously supplied with the drug. Subcutaneous implantation of sustained-release morphine pellets (75 mg) into the mice promotes the development of tolerance and also allows the daily measurement of antinociception. This approach is a powerful means for assessing both tolerance and dependence in rodents^{10–13}. Although wild-type littermates experienced a marked decline in responsiveness over 72 h of uninterrupted morphine, the $\beta\text{arr}2^{-/-}$ mice maintained consistent nociceptive latencies (Fig. 3). The parental mouse strains (129/SvJ and C57/BL6) also developed tolerance to morphine in each of the models of chronic treatment (daily injection or pellet implantation) in a manner similar to the wild-type littermates (data not shown). Thus, whether tolerance to morphine is induced acutely, after a single high-dose injection, or chronically, by maintaining constant levels of the drug (morphine pellets) over several days, blocking μOR desensitization by inactivating β -arrestin-2 eliminates morphine antinociceptive tolerance.

Prolonged exposure to morphine evokes the development of physical dependence to the drug. Tolerance to morphine is often viewed as the gateway to the development of physical dependence¹⁴. As the $\beta\text{arr}2^{-/-}$ mice do not become tolerant to morphine in this scheme, we determined whether they would become dependent on the drug. Opiate dependence was determined by assessing withdrawal responses after administration of the opiate antagonist, naloxone, after 72 h of pellet implantation^{10–12}. Notably, although the $\beta\text{arr}2^{-/-}$ mice did not develop morphine tolerance, they still experienced withdrawal after administration of naloxone (0.04, 0.2 or 1.0 mg per kg, intraperitoneally (i.p.)), which indicates that these mice, like their wild-type littermates, had become dependent on morphine during the 72-h treatment (Fig. 4). There were no significant differences discernable between the genotypes at any of the doses of naloxone.

The cellular mechanisms that lead to opiate dependence are mostly unknown, but the most observed correlative biochemical adaptation is an upregulation of the adenosine-3',5'-monophosphate (cyclic AMP) pathway, most notably in the upregulation of adenylyl cyclases^{1–3,14–16}. Therefore, we evaluated adenylyl cyclase activity in striata from mice treated in the daily chronic morphine scheme (mice in which [³⁵S]GTP γ S binding and receptor levels were

assessed in brainstem membranes; see Fig. 2). After 5 d of chronic treatment, there was a significant increase in basal and forskolin-stimulated adenylyl cyclase activity in both the wild-type and $\beta\text{arr}2^{-/-}$ mice, as compared with saline-treated controls (Fig. 5). The observed increase in adenylyl cyclase activity, a hallmark of opiate dependence, in the $\beta\text{arr}2^{-/-}$ mice indicates that these mice may experience biochemical adaptations in the chronic presence of morphine similar to those occurring in wild type. Although the $\beta\text{arr}2^{-/-}$ mice do not become tolerant to morphine, they clearly become dependent on it, as assessed both behaviourally and biochemically.

Chronic opiate treatment results in functional uncoupling of the μOR and its G protein in both cell culture^{17–19} and animal models^{14,20,21}. Although studies in transfected cell cultures have implicated phosphorylation by GRKs (G-protein-coupled receptor kinases) and binding of β -arrestins in μOR desensitization, the contribution of these regulatory elements to the physiological regulation of the μOR ^{22–24}, particularly by morphine, and to the development of tolerance has not been previously tested directly *in vivo*. The use of the $\beta\text{arr}2^{-/-}$ mouse, in which the desensitization process is impaired by a specific genetic manipulation, provides direct evidence in an animal model for the essential contribution of the β -arrestin-2 molecule to the desensitization of the receptor, and to the development of tolerance, after chronic exposure to morphine. Besides the desensitization of the receptor, several downstream adaptive responses to chronic morphine treatment have been observed including increased expression of the transcription factors, CREB and ΔFosB , in certain brain regions and most prominently, an upregulation of the cAMP pathway^{1–3,14,25}. Loss of β -arrestin-2 does not interfere with this cellular aspect of chronic morphine exposure, as both genotypes experience withdrawal and exhibit an increase in adenylyl cyclase activity.

Our results shed light on the relationship between opiate receptor desensitization in response to an agonist such as morphine and the development of both tolerance to and physical dependence on the drug. We have shown that abrogating μOR desensitization abolishes the development of antinociceptive tolerance to acute or chronic morphine. Moreover, our findings—that in the absence of tolerance to morphine physical dependence persists—indicate that these two phenomena are dissociable, and that their underlying biochemical mechanisms are different. □

Methods

Mutant mice

Wild-type and knockout mice were derived from crossing (over eight generations) heterozygous $\beta\text{arr}2$ C57BL6/129SvJ animals as described⁶. $\beta\text{arr}2^{+/+}$ and $\beta\text{arr}2^{-/-}$ mice used in this study were age-matched, 3–5-month-old male siblings weighing between 20 and 30 g. In all experiments, wild-type littermates served as controls for the $\beta\text{arr}2^{-/-}$ mice and both genotypes were evaluated simultaneously. Experiments were conducted in accordance with the NIH guidelines for the care and use of animals and with an approved animal protocol from the Duke University Animal Care and Use Committee.

Acute tolerance

Acute antinociceptive tolerance was induced as described²⁶. In each antinociception assay, nociceptive latencies were assessed as the response time to the hot plate (56 °C). The 'response' was defined by the animal either licking his fore- or hindpaws or flicking his hindpaws. In these studies, the most prominent response was forepaw licking. To avoid tissue damage, an artificial maximum time for exposure was imposed, which prevented the animal from contact with the plate for greater than 30 s. Data are reported as the maximum possible effect (% MPE), which was determined by accounting for each individual's basal response as well as the imposed maximum cutoff time using the following calculation: $100\% \times [(\text{drug response time} - \text{basal response time}) / (30 \text{ s} - \text{basal response time})] = \% \text{ MPE}$. Twenty-four hours after the 100 mg per kg morphine treatment, both genotypes had returned to their basal nociceptive latencies.

Chronic daily morphine treatment

In this study, two groups of mice were treated. The first group, containing wild-type ($n = 6$) and $\beta\text{arr}2^{-/-}$ ($n = 6$) mice were treated daily (between 15:00 and 16:00) with morphine (10 mg per kg, s.c.) and antinociception was assessed 30 min later on each day for the first 5 d. The second group, containing wild-type ($n = 7$) and $\beta\text{arr}2^{-/-}$ ($n = 6$) mice were treated daily (between 15:00 and 16:00) with morphine (10 mg per kg, s.c.), and

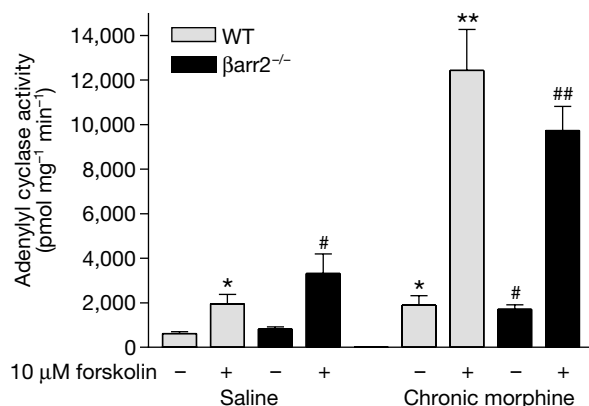


Figure 5 Adenylyl cyclase activity in WT and $\beta\text{arr}2^{-/-}$ mice after chronic morphine treatment. Adenylyl cyclase activity was assessed in membranes prepared from striata of WT and $\beta\text{arr}2^{-/-}$ mice after 5 d of chronic morphine treatment (10 mg per kg, s.c., per day). Data represent the mean $\pm$ s.e.m. of experiments performed in four striata from four animals of each genotype assayed in triplicate in parallel experiments. Asterisk, $P < 0.05$ versus untreated WT; two asterisks, $P < 0.01$ versus all WT; hash, $P < 0.05$ versus untreated $\beta\text{arr}2^{-/-}$; two hashes, $P < 0.02$ versus all $\beta\text{arr}2^{-/-}$; Student's *t*-test. There were no significant differences between the genotypes under the same treatment conditions.

antinociception was assessed 30 min after the injections on every odd day for 9 d. Dose-dependent responses were determined in the second group on day 1 and on day 9 using a cumulative dosing model^{6,27}.

³⁵S]GTPγS binding

Brain regions were dissected and immediately frozen in liquid nitrogen and were stored at -80 °C for less than 1 week before use. [³⁵S]GTPγS (1,250 Ci mmol, NEN, Boston, MA) binding was assessed as described^{6,28}. Briefly, samples were placed on ice and homogenized by polytron in membrane preparation buffer (50 mM Tris pH 7.4, 1 mM EDTA, 3 mM MgCl₂) and crude membranes were prepared by centrifugation at 20,000g for 15 min at 4 °C. Membranes were washed and pelleted twice in this buffer. Membranes were resuspended in assay buffer (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 3 mM MgCl₂, 0.2 mM EDTA) containing 10 μM GDP, and the samples (20 μg protein per tube) were incubated in 50 pM [³⁵S]GTPγS. [³⁵S]GTPγS binding to isolated brainstem membranes was determined after a 2-h stimulation (30 °C) with 50–10,000 nM DAMGO. Reactions were terminated by rapid filtration over GF/B filters (Brandel, Gaithersburg, MD) using a Brandel cell harvester. Filters were washed three times with ice-cold 10 mM Tris-HCl pH 7.4 and then counted in a liquid scintillation counter. The per cent stimulated [³⁵S]GTPγS binding was calculated by dividing unstimulated [³⁵S]GTPγS binding into each agonist-stimulated point.

Morphine pellet implantation

Sustained-release morphine pellets (75 mg) and comparable placebo pellets were provided by the National Institute on Drug Abuse. Pellets were implanted subcutaneously under chloral hydrate anaesthesia and secured by sutures. Mice were kept unrestrained in cages and provided freely with food and water. Nociceptive latencies were measured 4, 24, 48 and 72 h after the surgery.

Adenylyl cyclase activity

Adenylyl cyclase activity was assessed by measuring conversion of [³²P]ATP (31 Ci mmol⁻¹; NEN) to [³²P]cAMP by column elution following forskolin stimulation. Forskolin-stimulated adenylyl cyclase activity was determined in striata taken from mice that had been treated daily with saline or morphine for 5 d. Samples were homogenized by polytron in membrane preparation buffer (10 mM Tris pH 7.4, 5 mM EDTA) and crude membranes were prepared by centrifugation at 20,000g for 15 min at 4 °C. Membranes were washed and pelleted twice in this buffer and then were resuspended in assay buffer (75 mM Tris-HCl pH 7.4, 15 mM MgCl₂, 2 mM EDTA, 500 μM IBMX). Membrane protein aliquots (10 μg per tube) were assayed for 10 μM forskolin-stimulated adenylyl cyclase activity in a final volume of 50 μl, for 5 min at 25 °C as described^{29,30}.

Received 12 July; accepted 3 October 2000.

- Nestler, E. J. Under siege: The brain on opiates. *Neuron* **16**, 897–900 (1996).
- Nestler, E. J. & Aghajanian, G. K. Molecular and cellular basis of addiction. *Science* **278**, 58–63 (1997).
- Koob, G. F., Sanna, P. P. & Bloom, F. E. Neuroscience of addiction. *Neuron* **21**, 467–476 (1998).
- Ferguson, S. S., Barak, L. S., Zhang, J. & Caron, M. G. G-protein-coupled receptor regulation: role of G-protein-coupled receptor kinases and arrestins. *Can. J. Physiol. Pharmacol.* **74**, 1095–1110 (1996).
- Pitcher, J. A., Freedman, N. J. & Lefkowitz, R. J. G protein-coupled receptor kinases. *Annu. Rev. Biochem.* **67**, 653–692 (1998).
- Bohn, L. M. *et al.* Enhanced morphine analgesia in mice lacking β-arrestin 2. *Science* **286**, 2495–2498 (1999).
- Matthes, H. W. *et al.* Loss of morphine-induced analgesia, reward effect and withdrawal symptoms in mice lacking the μ-opioid-receptor gene. *Nature* **383**, 819–823 (1996).
- Sora, I. *et al.* Opiate receptor knockout mice define μ receptor roles in endogenous nociceptive responses and morphine-induced analgesia. *Proc. Natl Acad. Sci. USA* **94**, 1544–1549 (1997).
- Kieffer, B. L. Opioids: first lessons from knockout mice. *Trends Pharmacol. Sci.* **20**, 19–26 (1999).
- Way, E. L., Loh, H. H. & Shen, F. H. Simultaneous quantitative assessment of morphine tolerance and physical dependence. *J. Pharmacol. Exp. Ther.* **167**, 1–8 (1969).
- Brase, D. A., Loh, H. H. & Way, E. L. Comparison of the effects of morphine on locomotor activity, analgesia and primary and protracted physical dependence in six mouse strains. *J. Pharmacol. Exp. Ther.* **201**, 368–374 (1977).
- Schulteis, G., Markou, A., Gold, L. H., Stinus, L. & Koob, G. F. Relative sensitivity to naloxone of multiple indices of opiate withdrawal: a quantitative dose–response analysis. *J. Pharmacol. Exp. Ther.* **271**, 1391–1398 (1994).
- Jolas, T., Nestler, E. J. & Aghajanian, G. K. Chronic morphine increases GABA tone on serotonergic neurons of the dorsal raphe nucleus: association with an up-regulation of the cyclic AMP pathway. *Neuroscience* **95**, 433–443 (2000).
- Cox, B. M. in *Opioids in Pain Control: Basic and Clinical Aspects* (ed. C. Stein) 109–130 (Cambridge Univ. Press, New York, 1999).
- Collier, H. O. Cellular site of opiate dependence. *Nature* **283**, 625–629 (1980).
- Avdler-Reiss, T. *et al.* Adenylyl cyclase supersensitization in μ-opioid receptor-transfected Chinese hamster ovary cells following chronic opioid treatment. *J. Biol. Chem.* **270**, 29732–29738 (1995).
- Puttfarcken, P. S., Werling, L. L. & Cox, B. M. Effects of chronic morphine exposure on opioid inhibition of adenylyl cyclase in 7315c cell membranes: a useful model for the study of tolerance at μ-opioid receptors. *Mol. Pharmacol.* **33**, 520–527 (1988).
- Breivogel, C. S., Selley, D. E. & Childers, S. R. Acute and chronic effects of opioids on δ and μ receptor activation of G proteins in NG108-15 and SK-N-SH cell membranes. *J. Neurochem.* **68**, 1462–1472 (1997).
- Elliott, J., Guo, L. & Traynor, J. R. Tolerance to μ-opioid agonists in human neuroblastoma SH-SY5Y cells as determined by changes in guanosine-5'-O-(3-[35S]-thio)triphosphate binding. *Br. J. Pharmacol.* **121**, 1422–1428 (1997).
- Sim, L. J., Hampson, R. E., Deadwyler, S. A. & Childers, S. R. Effects of chronic treatment with Δ⁹-tetrahydrocannabinol on cannabinoid-stimulated [³⁵S]GTPγS autoradiography in rat brain. *J. Neurosci.* **16**, 8057–8066 (1996).

- Noble, F. & Cox, B. M. Differential desensitization of μ- and δ-opioid receptors in selected neural pathways following chronic morphine treatment. *Br. J. Pharmacol.* **117**, 161–169 (1996).
- Kovoor, A., Nappey, V., Kieffer, B. L. & Chavkin, C. μ and δ opioid receptors are differentially desensitized by the coexpression of β-adrenergic receptor kinase 2 and β-arrestin 2 in xenopus oocytes. *J. Biol. Chem.* **272**, 27605–27611 (1997).
- Zhang, J. *et al.* Role for G protein-coupled receptor kinase in agonist-specific regulation of μ-opioid receptor responsiveness. *Proc. Natl Acad. Sci. USA* **95**, 7157–7162 (1998).
- Whistler, J. L. & von Zastrow, M. Morphine-activated opioid receptors elude desensitization by β-arrestin. *Proc. Natl Acad. Sci. USA* **95**, 9914–9919 (1998).
- Lane-Ladd, S. B. *et al.* CREB (cAMP response element-binding protein) in the locus coeruleus: biochemical, physiological, and behavioral evidence for a role in opiate dependence. *J. Neurosci.* **17**, 7890–7901 (1997).
- Fairbanks, C. A. & Wilcox, G. L. Spinal antinociceptive synergism between morphine and clonidine persists in mice made acutely or chronically tolerant to morphine. *J. Pharmacol. Exp. Ther.* **288**, 1107–1116 (1999).
- Zhu, Y. *et al.* Retention of supraspinal delta-like analgesia and loss of morphine tolerance in delta opioid receptor knockout mice. *Neuron* **24**, 243–252 (1999).
- Gainetdinov, R. R. *et al.* Muscarinic supersensitivity and impaired receptor desensitization in G protein-coupled receptor kinase 5-deficient mice. *Neuron* **24**, 1029–1036 (1999).
- Hausdorff, W. P., Hnatowich, M., O'Dowd, B. F., Caron, M. G. & Lefkowitz, R. J. A mutation of the β₂-adrenergic receptor impairs agonist activation of adenylyl cyclase without affecting high affinity agonist binding. Distinct molecular determinants of the receptor are involved in physical coupling to and functional activation of Gs. *J. Biol. Chem.* **265**, 1388–1393 (1990).
- Zhang, J., Barak, L. S., Winkler, K. E., Caron, M. G. & Ferguson, S. S. A central role for β-arrestins and clathrin-coated vesicle-mediated endocytosis in β₂-adrenergic receptor resensitization. Differential regulation of receptor resensitization in two distinct cell types. *J. Biol. Chem.* **272**, 27005–27014 (1997).

Acknowledgements

We thank S. Suter for care and genotyping of animals. R.R.G. is a visiting scientist from the Institute of Pharmacology, Russian Academy of Medical Sciences, Baltiyskaya 8, 125315 Moscow, Russia. M.G.C. and R.J.L. are Investigators of the Howard Hughes Medical Institute.

Correspondence should be addressed to M.G.C. (e-mail: caron002@mc.duke.edu) or R.J.L. (e-mail: lefkow001@receptor-biol.duke.edu). Requests for materials should be addressed to R.J.L.

Tyrosine-kinase-dependent recruitment of RGS12 to the N-type calcium channel

Max L. Schiff*†, David P. Siderovski‡, J. Dedrick Jordan*, Greg Brothers\$, Bryan Snow\$, Luc De Vries||, Daniel F. Ortiz‡ & Maria Diversé-Pierluissi*

* Department of Pharmacology, Mount Sinai School of Medicine, New York, New York 10029, USA

‡ Department of Pharmacology, University of North Carolina, Chapel Hill, North Carolina 27559, USA

\$ Amgen Research Institute, Toronto M5G 2C1, Canada

|| Department of Cellular and Molecular Medicine, University of California-San Diego, La Jolla, California 92093, USA

‡ Department of Physiology, Tufts University School of Medicine, Boston, Massachusetts 02111, USA

† These authors contributed equally to this paper.

γ-Aminobutyric acid (GABA)_B receptors couple to G_o to inhibit N-type calcium channels in embryonic chick dorsal root ganglion neurons¹. The voltage-independent inhibition, mediated by means of a tyrosine-kinase pathway², is transient and lasts up to 100 seconds. Inhibition of endogenous RGS12, a member of the family of regulators of G-protein signalling, selectively alters the time course of voltage-independent inhibition. The RGS12 protein, in addition to the RGS domain, contains PDZ and PTB domains³. Fusion proteins containing the PTB domain of RGS12 alter the rate of termination of the GABA_B signal, whereas the PDZ or RGS domains of RGS12 have no observable effects. Using

primary dorsal root ganglion neurons in culture, here we show an endogenous agonist-induced tyrosine-kinase-dependent complex of RGS12 and the calcium channel. These results indicate that RGS12 is a multifunctional protein capable of direct interactions through its PTB domain with the tyrosine-phosphorylated calcium channel. Recruitment of RGS proteins to G-protein effectors may represent an additional mechanism for signal termination in G-protein-coupled pathways.

Activation of GABA_B receptors in chick dorsal root ganglion neurons inhibits the N-type calcium channel in both a voltage-dependent and a voltage-independent manner (Fig. 1a). The termination of transmitter-mediated modulation of the N-type calcium current requires the activation of a G-protein-coupled receptor kinase, GRK3 (ref. 4). Despite this common requirement for desensitization, G_i- and G_o-mediated pathways show different rates of desensitization⁴. We have previously shown that these differences may be due to the selective effects of regulators of G-protein signalling, RGS proteins⁵.

To study the role of RGS proteins in this desensitization, we introduced recombinant human RGS12 (1.0 nM) into the dorsal root ganglion neuron cell bodies by micro-injection. After 30 min, we measured calcium current before and during continuous exposure to 100 μ M GABA. RGS12 selectively accelerated the rate of reversal of the voltage-independent pathway mediated by tyrosine kinase, whereas the voltage-dependent pathway remained unaltered (Fig. 1b). The half-time value for termination in control cells was 42 ± 8 s compared with 2.4 ± 1 s in RGS12-treated cells. Experiments in which RGS12 (0.02–1.00 nM, 20–25 min equilibration time) was introduced through the recording pipette yielded similar results (data not shown). The magnitude of the GABA-mediated inhibition of the calcium current was unaltered (see Supplementary

Information). There was no change observed in the extent or rate of current rundown. Other RGS proteins, such as GAIP, RGS2 and RGS14, failed to alter the GABA-mediated response (Fig. 1c). Injection of recombinant RGS protein did not have an effect on the magnitude of the basal calcium current. A similar degree of selectivity was observed by using antibodies against different endogenous RGS proteins (see Supplementary Information).

We next determined the role of endogenous RGS12. A complementary DNA fragment showing 85% identity with the rat RGS12 nucleotide sequence was generated from chick sensory neuron RNA (GenBank accession no. AF090086). Antibodies directed against rat RGS12 recognize a protein that has a relative molecular mass (M_r) of 140,000 (140K) that is mainly associated with the membrane fraction and does not redistribute on agonist stimulation (Fig. 1d). Other RGS proteins, such as RGS4 and GAIP, have also been found in membrane-associated fractions^{6,7}. Cells that had been equilibrated with an antibody (immunoglobulin- γ) against the RGS box of RGS12 ($n = 7$) showed a half-time for desensitizing voltage-independent inhibition of 605 ± 14 s compared with 42 ± 8 s ($n = 10$) in control cells (Fig. 1e). Antibodies raised against the amino terminus of rat RGS12 also slowed the desensitization rate of the GABA-mediated voltage-independent inhibition of N-type calcium current (data not shown). Introduction of the antibodies did not alter the properties of the calcium current including the current-voltage relationship (see Supplementary Information). These results indicate that the endogenous RGS12 has a function in determining the kinetics of desensitization of GABA_B-mediated inhibition of the N-type calcium channel.

The N terminus of RGS12 contains a PDZ domain⁸ and a PTB domain³. Because the antibodies raised against the N terminus of RGS12 blocked desensitization, we tested the ability of fusion

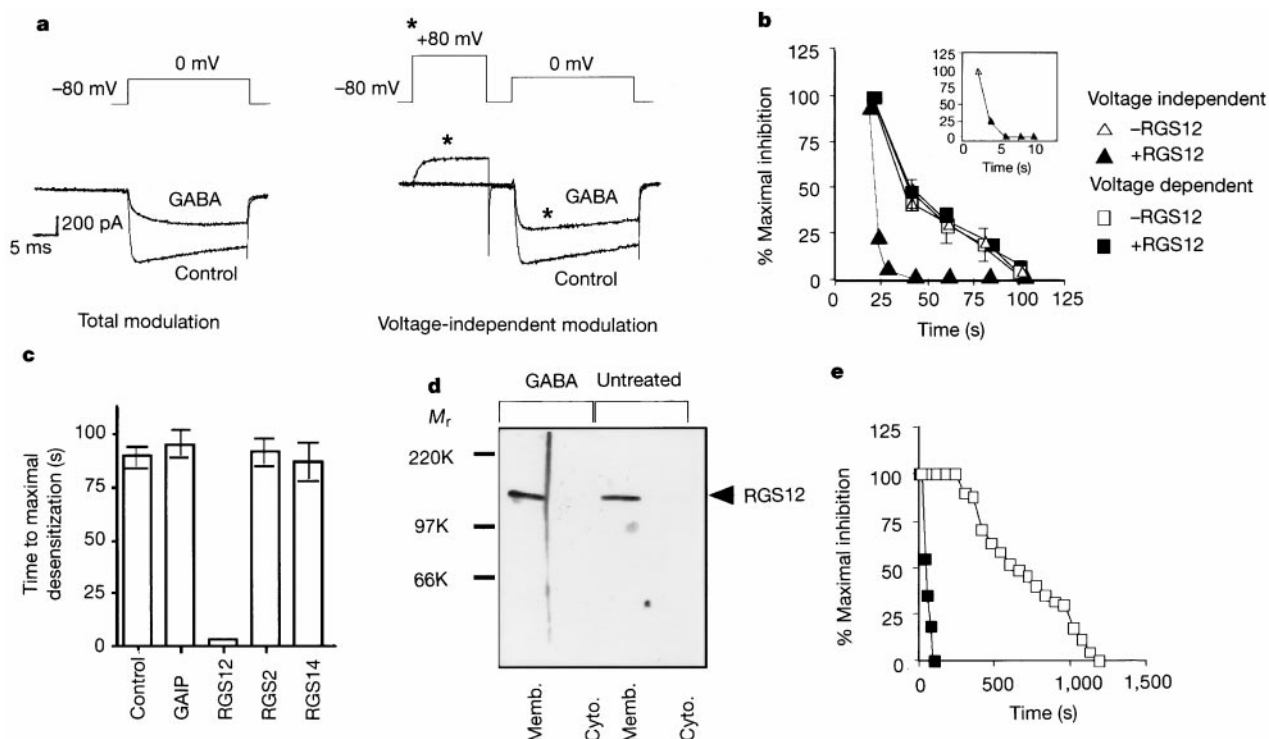


Figure 1 Effect of RGS12 on the desensitization of the N-type calcium channel. **a**, Separation of GABA-induced voltage-dependent from voltage-independent inhibition. Inhibition of calcium current by 100 μ M GABA was measured with (asterisk) or without a 30-ms prepulse to +80 mV. **b**, Effects of RGS12 on time course of desensitization of the GABA_B-mediated inhibition of calcium current. Time point '0' is time at which maximal inhibition was observed. Squares represent voltage-dependent inhibition and triangles represent voltage-independent inhibition. Open symbols represent control solution and

filled symbols represent presence of RGS12. Each data point is the mean of at least five cells. **c**, Time point at which maximal desensitization was achieved in presence of different RGS proteins. **d**, Association of RGS12 with the membrane. Immunodetection using antibodies raised against the RGS12 box. **e**, Effect of an anti-RGS12 antibody on time course of termination of the GABA_B-mediated response. Open squares represent cells injected with the antibodies and filled squares represent control cells.

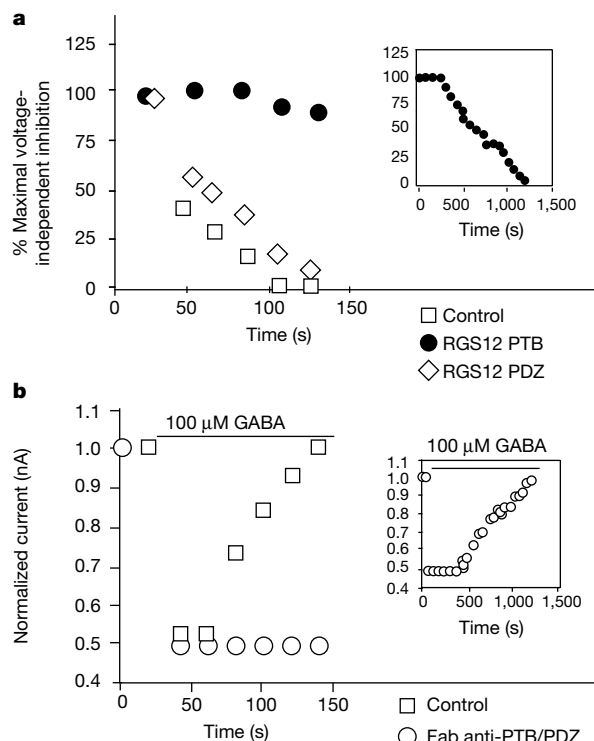


Figure 2 Effects of two domains of RGS12 on rate of desensitization of GABA_B-mediated inhibition of calcium current and the Fab fragment on the time course of termination of the GABA_B-mediated response. **a**, Effects of the PTB and PDZ domains of RGS12 on desensitization rate of GABA_B-mediated inhibition of calcium current. Percentage of inhibition as a function of time was plotted before and during transmitter application. Time points represent mean values from five cells. Open squares represent control cells, open

diamonds represent cells exposed to human RGS12 PDZ domain and filled circles represent cells exposed to human RGS12 PTB domain. **b**, Effect of the Fab fragment of the anti-PDZ/PTB RGS12 antibody on time course of termination of the GABA_B-mediated response. Antibodies (1:1,000) were introduced into the cell bodies by micro-injection. Each time point is the mean of six cells.

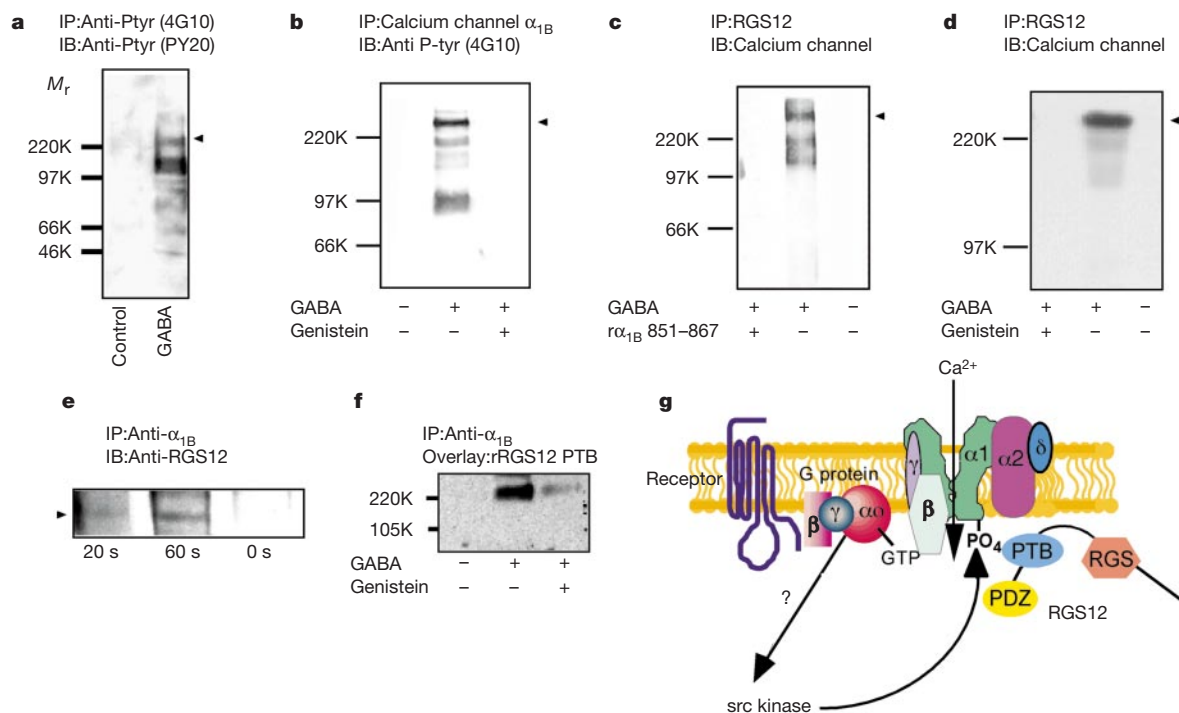


Figure 3 Formation of a regulated complex between RGS12 and the α_{1B} subunit of the calcium channel. **a**, Immunoblot (IB) to detect tyrosine-phosphorylated proteins following immunoprecipitation (IP) using anti-phosphotyrosine antibody (4G10). Immunodetection was performed using a second anti-phosphotyrosine antibody (PY20). **b**, Anti- α_{1B} antibody was used to precipitate the calcium channel and immunoblotting was performed using an anti-phosphotyrosine antibody (4G10). **c**, Precipitation of RGS12 followed by immunoblotting with antibody against the α -subunit of the calcium channel. **d**, Interaction

of RGS12 and the calcium channel is prevented by a tyrosine-kinase inhibitor. **e**, Interaction of RGS12 with the calcium channel is time dependent. Immunoblots were probed to detect RGS12 in samples from anti- α_{1B} precipitates. **f**, Overlay assays with PTB domains of rat RGS12. Anti- α_{1B} antibody was used to precipitate calcium channel from control and GABA-treated cells. Data are representative of three experiments. **g**, Interactions of a Src-like kinase with the calcium channel and RGS12. Interactions that we found in experiments are indicated.

proteins containing the N-terminal region (residues 1–440), or the PDZ (residues 1–99) or PTB (residues 236–440) domains of RGS12 to modulate the endogenous desensitization or desensitization in the presence of recombinant RGS12. Fusion proteins containing the PTB domain of RGS12 altered the rate of desensitization, whereas the PDZ domain had no effect on this rate (Fig. 2a; also see Supplementary Information). To determine whether the native RGS12 protein also used its PTB domain in modulating desensitization, we tested the effect of Fab fragments obtained from antibodies raised against the N-terminal PDZ/PTB domain of RGS12. The Fab fragments of anti-PTB/PDZ antibodies slowed the rate of desensitization of the voltage-independent inhibition of calcium current mediated by GABA, without altering the onset or magnitude of the response (Fig. 2b).

These results suggest that interaction of the PTB domain of RGS12 with tyrosine-phosphorylated proteins may be important for the timing of desensitization. There are several possible candidates, including the G proteins and the channel itself. We first identified a tyrosine-phosphorylated protein with an apparent M_r of 240K (Fig. 3a). Immunoprecipitation with the anti-tyrosine antibody did not show any G-protein subunits. The 240K size is similar to the reported size range of the α -subunit of the N-type channel (α_{1B}) of 170–260K (refs 9–11). We determined whether this 240K protein represents the chick α_{1B} calcium channel. Immunoblot analysis using the membrane fractions from chick DRG neurons shows that a protein of M_r 240K interacts with an antibody that recognizes loop II/III of the α_1 -subunit of the N-type calcium channel (see Supplementary Information). Similar bands to those observed in immunoprecipitation using the anti- α_{1B} antibody against the region that binds to synaptic vesicle proteins are obtained using another calcium-channel antibody raised against residues 1,382–1,400 of the rat α_1 -subunit from skeletal muscle (see Supplementary Information). This region is conserved across all of the α_1 -subunits of high-voltage-activated calcium channels. This pattern of bands using both of the antibodies indicates that there may be degradation of the channel protein during processing, although we did not explore this in detail.

When the anti- α_{1B} immunoprecipitate was blotted with an anti-phosphotyrosine antibody (4G10), a similar pattern to those in previous experiments² using 32 P was observed (Fig. 3b). Pretreatment of the cells with a tyrosine kinase inhibitor (genistein) blocked the receptor-dependent appearance of this band. From these results, we conclude that the 240K protein represents the α -subunit of the N-type calcium channel, and that this subunit is tyrosine-phosphorylated in response to GABA stimulation.

Because the α_{1B} subunit of the N-type calcium channel is tyrosine phosphorylated and as the PTB domain of RGS12 is involved in desensitization, we next determined whether the channel interacts with RGS12. We performed an immunoprecipitation with anti-RGS12 antibodies followed by immunoblotting with anti- α_{1B} antibody. The α_1 -subunit of the calcium channel co-precipitates with RGS12 in cells treated with GABA, whereas the complex could not be detected in untreated cells (Fig. 3c). Pre-incubation of the anti- α_{1B} antibody with the peptide used as immunogen (rat α_{1B} , residues 851–867) blocked cross-reactivity, showing that recognition of the protein is specific (Fig. 3c). Bands in the 97K–220K range are also blocked, suggesting that they may be fragments of the 240K protein.

We have previously reported that genistein prevents the voltage-independent inhibition mediated by GABA². In the presence of genistein, a tyrosine kinase inhibitor, the calcium channel did not co-precipitate with RGS12 (Fig. 3d). This suggests that tyrosine phosphorylation is needed for the interaction between RGS12 and the calcium channel. Together the experiments shown in Fig. 3 indicate that RGS12 and the N-type calcium channel form a complex in a tyrosine-kinase-dependent manner.

The interaction of RGS12 with the calcium channel is time dependent. At an early time point (20 s), when the inhibition is

still maximal, very little RGS12 co-precipitates with the calcium channel (Fig. 3e). During desensitization (Fig. 3e) more of the RGS12 co-precipitates with the channel. We next determined whether the interaction between RGS12 and the calcium channel is direct by using a protein overlay assay. The PTB domain of RGS12 (rRGS12, residues 193–394) binds to the α_1 -subunit of the calcium channel (Fig. 3f). This interaction is prevented by the tyrosine kinase inhibitor genistein, indicating that there may be direct interaction between RGS12 and the tyrosine-phosphorylated calcium channel.

These results indicate that tyrosine-kinase-mediated phosphorylation of a G-protein effector, such as the calcium channel, is involved in the termination of G-protein-mediated signalling by facilitating recruitment of a signal termination molecule to the effector. Thus tyrosine kinases may act both as 'on' and 'off' molecular switches in the G-protein pathway. In this pathway a Src-like kinase is involved in both the onset and the termination of the response. During the onset of the response, this tyrosine kinase phosphorylates the α_1 -subunit of the N-type channel. The tyrosine-phosphorylated form of the channel α_1 -subunit becomes a potential target for proteins that contain a PTB domain such as RGS12. These interactions are summarized in Fig. 3g. RGS12 may act to terminate the signal by two mechanisms: RGS12 may directly modulate channel activity; alternatively, the channel may be used as a 'scaffold' to recruit RGS12 to the vicinity of $G\alpha_2$ and functional effects may arise from RGS12 enhancing the GTPase activity of $G\alpha$. Whether either or both of these mechanisms are in operation remains to be determined. Nevertheless, the many domains of RGS proteins endow them with varied functional capabilities in addition to regulation of GTPase activity, thus making them versatile regulators of signalling that can interact with receptors¹², G proteins (for review see ref. 7) and now effectors. □

Methods

Cell culture

We grew embryonic chick sensory neurons in culture as described¹³. We dissected dorsal root ganglia from 11–12-day-old embryos. Cells to be used for electrophysiology were plated at a density of about 50,000 cells per 35-mm collagen-coated tissue-culture dish, and studied between 1 and 3 days *in vitro*.

Antibodies

Antisera directed against the N-terminal PDZ–PTB tandem (residues 1–440) and RGS box (residues 664–885) of rat RGS12 were generated and affinity purified as described⁸. Affinity-purified anti- α_{1B} calcium-channel antibody was purchased from Alomone Laboratories (ACC-002), and monoclonal antibodies against phosphotyrosine PY20 and 4G10 were purchased from Upstate Biotechnology. Control mouse and rabbit IgG were purchased from Pierce. We introduced antibodies (1:1,000) into the cell bodies by microinjection.

Electrophysiology

We performed whole-cell recordings as described¹³. For extracellular application, we diluted agents into standard extracellular saline and applied them through a wide-bore pipette. In all experiments, calcium current has been corrected for rundown by measuring calcium current as a function of time in control cells without transmitter. Cells used for experiments showed a rundown of the current of less than 1% per minute.

Fusion proteins

Recombinant GST-tagged human RGS12 PDZ domain (residues 1–110), rat RGS12 PDZ domain (residues 1–94) and thioredoxin-tagged rat RGS12 PDZ–PTB tandem (residues 1–440) proteins were expressed and purified as described⁸. We introduced fusion proteins (200 ng ml⁻¹) into the cells by passive diffusion through the recording pipette. After an equilibration period of 15 min we applied the transmitter.

Data analysis

Data were filtered at 3 kHz, acquired at 10–20 kHz and analysed using PulseFit (HEKA) and IgorPro (WaveMetrics) on a Macintosh G3 computer. Strong depolarizing conditioning pulses (to 80 mV) that precede test pulses (to 0 mV) reverse voltage-dependent inhibition induced by GABA without affecting voltage-independent inhibition. Such conditioning pulses have no effect on control currents recorded in the absence of GABA. During the application of the transmitter, test-pulse currents measured before and after the conditioning pulse are subtracted to yield the voltage-dependent component. Test

pulses measured following the conditioning pulse are subtracted from control currents (measured in the absence of GABA) to yield the voltage-independent component. Integration (total time) of these two currents gives measurements of total-charge entry carried by the two modulatory components.

Preparation of membrane fractions

For each condition we used 10^6 cells. Cells were disrupted using 50 mM Tris, pH 7.4, containing 250 μ M sodium pervanadate, 1 mM Pefabloc, 1 mM EDTA, 1 mM EGTA, 10 μ g ml⁻¹ pepstatin, 10 μ g ml⁻¹ leupeptin, 100 μ g ml⁻¹ soybean trypsin inhibitor and 100 μ g ml⁻¹ of each calpain I inhibitor and calpain II inhibitor. We centrifuged the homogenates at 1,000g for 10 min. Supernatants were recentrifuged at 100,000g for 1 h and the pellet was resuspended in ice-cold buffer (phosphate-buffered saline, pH 7.4, containing 250 μ M sodium pervanadate, 1% (v/v) NP-40, 1 mM Pefabloc, 1 mM EDTA, 1 mM EGTA, 10 μ g ml⁻¹ pepstatin, 10 μ g ml⁻¹ leupeptin, 100 μ g ml⁻¹ soybean trypsin inhibitor, 100 μ g ml⁻¹ calpain I inhibitor and 100 μ g ml⁻¹ calpain II inhibitor). The cytosolic and membrane-associated fractions were resolved on 6% SDS-PAGE. We detected RGS12 by immunoblotting using a 1:2,000 dilution of the antibodies.

Immunoprecipitation

We stimulated cells for 42 s with control solution containing 100 μ M bicuculline (inhibitor of GABA_A receptors) or 100 μ M GABA in the presence of bicuculline. For immunoprecipitation of tyrosine-phosphorylated proteins, we used 5×10^5 cells for each experimental condition. For immunoprecipitation of calcium channel α_{1B} -subunit and RGS12, 10^6 cells were used. Cells were exposed for 45 s to buffer $\pm$ 100 μ M GABA (both containing 100 μ M bicuculline to block GABA_A receptors) and lysed with ice-cold buffer (phosphate-buffered saline, pH 7.4, containing 250 μ M sodium pervanadate, 1% (v/v) NP-40, 1 mM Pefabloc, 1 mM EDTA, 1 mM EGTA, 10 μ g ml⁻¹ pepstatin, 10 μ g ml⁻¹ leupeptin, 100 μ g ml⁻¹ soybean trypsin inhibitor, 100 μ g ml⁻¹ calpain I inhibitor and 100 μ g ml⁻¹ calpain II inhibitor). Lysates were centrifuged at 15,000g for 30 min, and supernatant removed and exposed to untreated Protein A/G beads (Santa Cruz). The cleared preparation was incubated in primary antibody (6 μ g ml⁻¹ mouse monoclonal antibody against phosphotyrosine, 1:200 for the calcium channel and 1:2,000 for other proteins) for 16 h at 4 °C followed by incubation (4 h at 4 °C) in Protein A/G bead complex (Santa Cruz). Precipitates were collected by centrifugation and washed three times in ice-cold phosphate-buffered saline (250 μ M sodium pervanadate, 1% bovine serum albumin, 1 mM Pefabloc, 10 μ g ml⁻¹ pepstatin, 10 μ g ml⁻¹ leupeptin, and 100 μ g ml⁻¹ soybean trypsin inhibitor).

Received 3 May; accepted 22 August 2000.

1. Luebke, J. I. & Dunlap, K. Sensory neurons N-type calcium currents are inhibited by both voltage-dependent and -independent mechanisms. *Pflügers Archiv* **428**, 499–507 (1994).
2. Diversé-Pierluissi, M. A., Remmers, A. E., Neubig, R. & Dunlap, K. Novel form of crosstalk between G protein and tyrosine kinase pathways. *Proc. Natl Acad. Sci. USA* **94**, 5417–5421 (1997).
3. De Vries, L. & Farquhar, M. G. RGS proteins: more than just GAPs for heterotrimeric G proteins. *Trends Cell Biol.* **9**, 138–144 (1999).
4. Diversé-Pierluissi, M., Inglese, J., Stoffel, R. H., Lefkowitz, R. J. & Dunlap, K. G protein-coupled receptor kinase mediates desensitization of norepinephrine-induced calcium current inhibition. *Neuron* **16**, 579–585 (1996).
5. Diversé-Pierluissi, M. et al. RGS proteins as determinants of the rate of desensitization of neuronal calcium channels. *J. Biol. Chem.* **274**, 14490–14494 (1999).
6. De Vries, L. et al. RGS-GAIP, a GTPase-activating protein for G α i heterotrimeric G proteins, is located on clathrin-coated vesicles. *Mol. Biol. Cell.* **9**, 1123–1134 (1998).
7. Srinivasa, S. P., Bernstein, L. S., Blumer, K. J. & Linder, M. E. Plasma membrane localization is required for RGS4 function in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **95**, 5584–5589 (1998).
8. Snow, B. E. et al. GTPase activating specificity of RGS12 and binding specificity of an alternatively spliced PDZ (PSD-95/Dlg/ZO-1) domain. *J. Biol. Chem.* **273**, 17749–17755 (1998).
9. Dubel, S. J. et al. Molecular cloning of the α -1 subunit of an omega-conotoxin-sensitive calcium channel. *Proc. Natl Acad. Sci. USA* **89**, 5058–5062 (1992).
10. Ahljianian, M. K., Striessnig, J. & Catterall, W. A. Phosphorylation of an α 1-like subunit of an omega-conotoxin-sensitive brain calcium channel by cAMP-dependent protein kinase and protein kinase C. *J. Biol. Chem.* **266**, 20192–20197 (1991).
11. McEnery, M. W., Snowman, A. M., Sharp, A. H., Adams, M. E. & Snyder, S. H. Purified omega-conotoxin GVIA receptor of rat brain resembles a dihydropyridine-sensitive L-type calcium channel. *Proc. Natl Acad. Sci. USA* **88**, 11095–11099 (1991).
12. Zeng, W. et al. The N-terminal domain of RGS4 confers receptor-selective inhibition of G protein signaling. *J. Biol. Chem.* **273**, 34687–34690 (1998).
13. Diversé-Pierluissi, M., Goldsmith, P. K. & Dunlap, K. Transmitter-mediated inhibition of N-type calcium current channels in sensory neurons involves multiple GTP-binding proteins and subunits. *Neuron* **14**, 191–200 (1994).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank R. Iyengar, E. Landau and R. Blitzer for reading the manuscript. This work was supported by grants from the National Institutes of Health (NINDS), New York City Speaker's Fund for Biomedical Research and Mount Sinai Dean's Incentive Fund for Innovative Research to M.D.P.

Correspondence and requests for materials should be addressed to M.D.P. (e-mail: diverm01@doc.mssm.edu).

Adherens junctions and β -catenin-mediated cell signalling in a non-metazoan organism

Mark J. Grimson*†, Juliet C. Coates†‡§, Jonathan P. Reynolds‡, Mark Shipman‡, Richard L. Blanton* & Adrian J. Harwood‡

* Dept of Biological Sciences, Texas Tech University, Lubbock, Texas 79409, USA

‡ MRC Laboratory for Molecular Cell Biology, University College London, Gower St, London WC1E 6BT, UK

† These authors contributed equally to this work.

Mechanical forces between cells have a principal role in the organization of animal tissues. Adherens junctions are an important component of these tissues, connecting cells through their actin cytoskeleton and allowing the assembly of tensile structures^{1–4}. At least one adherens junction protein, β -catenin, also acts as a signalling molecule, directly regulating gene expression^{5–7}. To date, adherens junctions have only been detected in metazoa, and therefore we looked for them outside the animal kingdom to examine their evolutionary origins. The non-metazoan *Dictyostelium discoideum* forms a multicellular, differentiated structure⁸. Here we describe the discovery of actin-associated intercellular junctions in *Dictyostelium*. We have isolated a gene encoding a β -catenin homologue, *aardvark*, which is a component of the junctional complex, and, independently, is required for cell signalling. Our discovery of adherens junctions outside the animal kingdom shows that the dual role of β -catenin in cell–cell adhesion and cell signalling evolved before the origins of metazoa.

Dictyostelium discoideum grows as a unicellular amoeba, but when starved undergoes multicellular development. The resulting fruiting body is a differentiated structure of up to 10^5 cells that comprises a spore head supported by a single, vertical stalk. The stalk forms by elongation through the addition of differentiated stalk cells to its top. This process involves coordination of cell differentiation, cell movement and cell shape change. By using an ultra-rapid cryopreservation method and electron microscopy, we observed an unexpected complexity in the developing fruiting body (Fig. 1).

Longitudinal sections of the developing fruiting body show a constriction near to the top of the stalk tube (Fig. 1a). Transverse sections across this area show a ring of cells with close contacts or junctions between their lateral membranes (Fig. 1b–d). Each region of contact contains electron-dense material adjacent to the plasma membrane. Bundles, roughly 600 nm in diameter, of fine intracellular filaments, connect the junction on one side of the cell to that on the other (Fig. 1c, d). The filament bundles stain with anti-actin antibody (data not shown), with over 80% of the gold particles overlaying the filaments. The developing stalk tube is therefore enclosed, and likely to be compressed, within a ring of actin. Cells containing junctions are only found in the area of the constriction, and no junctions are observed in longitudinal sections. This suggests that cell junctions are only present in rings of cells perpendicular to the stalk tube axis, and that the constriction comprises a stack of such cellular rings. There are likely to be up to 10 rings, representing a total of 700–900 cells.

To visualize this structure in living cells, we expressed a protein fusion of green fluorescent protein (GFP) and the highly conserved actin-binding domain (ABPD) (relative molecular mass 25,000) from the amino terminus of the actin crosslinking protein ABP-120

§ Present address: MRC, Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, UK

(ref. 9); this protein binds exclusively to F-actin. In the developing fruiting body observed under low magnification, ABPD–GFP highlights a single population of cells that is dense in actin filaments: this coincides with the position of the constriction and encircles the stalk tube (Fig. 2; Supplementary Information). The morphology of the cell junctions and their association with actin filaments gives them the appearance of metazoan adherens junctions¹⁰.

A search of the *Dictyostelium* complementary DNA database¹¹ using the BLAST 2.0 algorithm¹² identified a potential β -catenin homologue that we called *aardvark* (*aar*; Fig. 3a). β -catenin contains a repetitive structural unit, known as the Armadillo (Arm) repeat, and the *aar* complementary DNA encodes a 758-amino-acid protein containing ten Arm repeats—two less than in metazoan β -catenin (Fig. 3b). The Arm repeats of Aar have roughly 50% similarity to those found in β -catenin. A comparable degree of similarity is seen between the *Caenorhabditis elegans* and vertebrate β -catenins. There are additional features in common between Aar and β -catenin protein sequences. Aar shows a potential α -catenin-binding site in the first Arm repeat of Aar, an equivalent position to the one present in β -catenin¹³ (Fig. 3b). Aar also contains a cluster of potential glycogen synthase kinase (GSK-3) phosphorylation sites in the amino terminus, as does β -catenin (Fig. 3b). Aar differs from metazoan β -catenin by having a truncated carboxy terminus, a structure confirmed by compiling the genomic sequence of *aar* (data not shown).

Aar also shows homology to the Arm repeats of a *Saccharomyces cerevisiae* protein Vac8, and sequences present in the *Schizosaccharomyces pombe* (SPBC354) and *Arabidopsis thaliana* (AB016888; encoded in P1 clone MDH9) genomic data bases. Vac8 is associated with the yeast vacuole and is required for vacuolar morphology and inheritance, and protein targeting to the vacuole; it may also associate with the actin cytoskeleton on the vacuolar membrane^{14–16}. Vac8, however, does not associate with the plasma membrane and there is no evidence for it possessing signalling functions. Furthermore, Vac8 does not contain an α -catenin-binding site or GSK-3 phosphorylation sites. Conversely, Aar and β -catenin proteins do not contain the consensus acylation sites essential for Vac8 localization and function^{14,16} (Fig. 3b). In growing cells, Aar protein is present throughout the cytoplasm (Fig. 3c). Mutants lacking *aar* show no defects in growth or division, suggesting that Aar is unlikely to be required for basic cell functions such as membrane integrity or trafficking (Fig. 3d). Finally, immuno-electron microscopy shows that the Aar protein is localized at the cell junctions in the developing fruiting body (Fig. 3e). These observations are all consistent with Aar being a β -catenin homologue.

We ablated the *aar* gene by gene disruption through homologous recombination (Fig. 4a). *aar* mutant cells develop, but form mechanically weak fruiting bodies, most of which collapse onto the substratum. The constriction of the stalk tube (Fig. 4b), the intercellular junctions (Fig. 4c) and the actin ring (Fig. 4d) are completely absent in the *aar* mutant. Expression of the *aar* cDNA restores adherens junctions to the *aar* mutant, and overexpression leads to the formation of ectopic junctional complexes in the pre-stalk cells that surround the developing stalk tube. These ectopic junctions cover an expanded area of the plasma membrane, causing it to become buckled (Fig. 4e). These observations show that Aar is required for *Dictyostelium* adherens junctions.

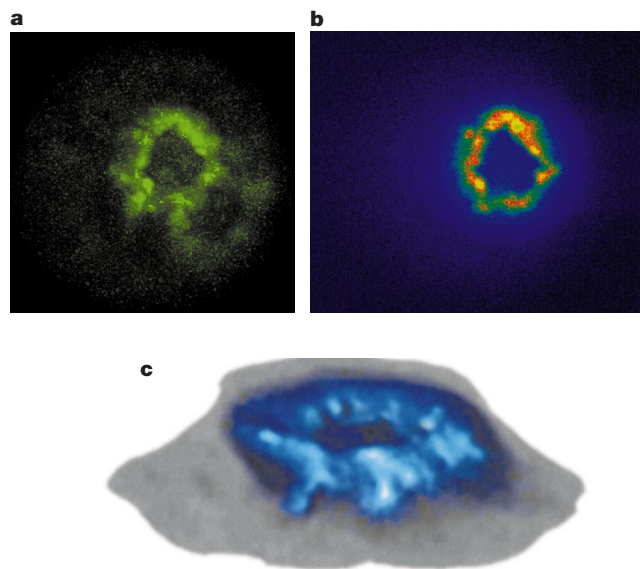


Figure 2 Direct visualization of the actin ring. Wild-type (AX2) cells expressing the ABPD–GFP fusion protein were developed to the preculminant stage (when the fruiting body forms), and low magnification confocal microscopy was used to collect optical sections from below the tip to the beginning of the spore head (see Fig. 1a). **a**, A three-dimensional (3D)-rendered image containing 70 0.5- μ m confocal sections viewed from above (original magnification $\times 20$). GFP is concentrated in a ring of cells around the stalk tube. **b**, A colour contour map of a single confocal image used to make **a**. Yellow, red and green correspond to cells enriched in F-actin and blue areas to low F-actin content. **c**, A 3D-rendered image of the F-actin ring visualized from an oblique angle. The ring of cells high in actin is coloured blue and the surrounding cells of the preculminant are coloured grey.

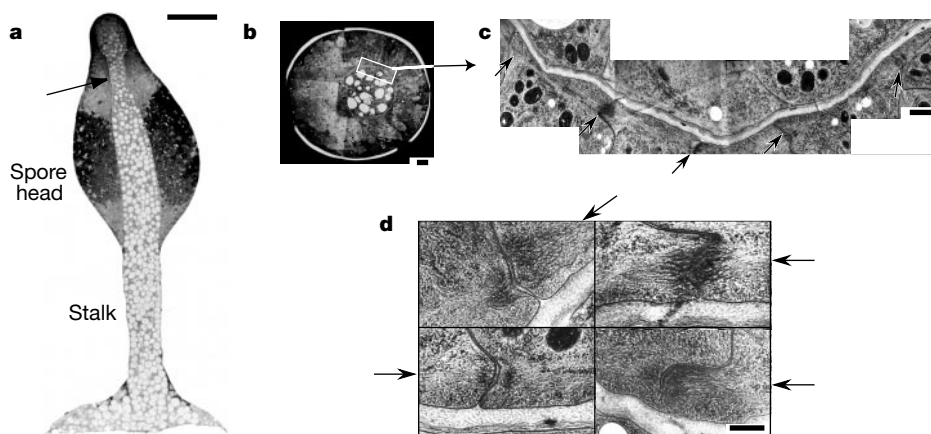


Figure 1 Identification of adherens junctions in the *Dictyostelium* fruiting body. **a**, Longitudinal 0.5- μ m section through a preculminant developed for 18 h and visualized by light microscopy. The constriction is marked by arrow. Scale bar, 50 μ m. **b**, Electron micrograph of a transverse section through the constriction. Scale bar, 50 μ m.

c, Magnification of region marked by box in **b**. Arrows show the position of cell junctions. Scale bar, 1 μ m. **d**, Montage of high magnification view of four junctions. Arrows mark the positions of 600-nm, actin filament bundles. Scale bar, 500 nm.

Dictyostelium spore and stalk cells are highly differentiated, but arise from precursors that are identified by their cell-type-specific gene expression¹⁷. In addition to its structural role, we find that Aar is required for spore-cell differentiation. Initially, we found that expression of *psA*, a prespore cell-specific marker, is reduced during

development of the *aar* mutant (Fig. 5a). In wild-type cells, *psA* is expressed at the mound stage of development, 8–10 h before the appearance of cell junctions. This reduced expression in the *aar* mutant is unlikely, therefore, to be a consequence of the failure to form adherens junctions. In addition, a number of additional experiments show that this effect is cell-autonomous. Spore differentiation was examined in low-density monolayer culture, where full differentiation is induced with 8-bromo-cAMP. No cell contact is observed between cells developed under these conditions, yet the *aar* mutant forms 50% fewer spores (Fig. 5b). The effect of Aar on gene expression is most pronounced when *psA* is induced by extracellular cAMP in isolated cells. Under these conditions, we find that expression of either a *lacZ* marker gene from the *psA* promoter (Fig. 5c) or the endogenous *psA* mRNA (Fig. 5d) is completely lost. Re-expression of the *aar* cDNA in the *aar* mutant restores *psA* gene expression (Fig. 5d). Furthermore, overexpression of the *aar* cDNA in wild-type cells causes increased *psA* gene expression (Fig. 5e). These results show that Aar is required for prespore gene expression.

In metazoa, β -catenin signalling involves its phosphorylation by GSK-3. Aar contains potential GSK-3 phosphorylation sites, and there is a strong reduction in the extracellular cAMP-mediated induction of *psA* in the absence of the *Dictyostelium* GSK-3 homologue *gskA*^{18,19}. We therefore examined the effects of overexpression of *aar* cDNA in a mutant that lacks all GskA activity. In contrast to the case in wild-type cells (Fig. 5d), overexpression of *aar* in a *gskA* mutant failed to elevate *psA* gene expression (Fig. 5f). This argues that the effect of Aar on gene expression requires GSK-3, and is consistent with a model in which GskA phosphorylation activates Aar signalling. These observations do not conform to the model proposed for β -catenin signalling in *Drosophila* and vertebrates²⁰; however, a similar situation is observed during early development of the nematode *C. elegans*. Here, both *gsk-3* and *wrm-1*, one of three *C. elegans* β -catenin homologues, act positively to specify an endodermal cell lineage^{21,22} (Fig. 5g).

The discovery of cell junctions and a possible β -catenin homologue in *Dictyostelium* leads to conclusions concerning their

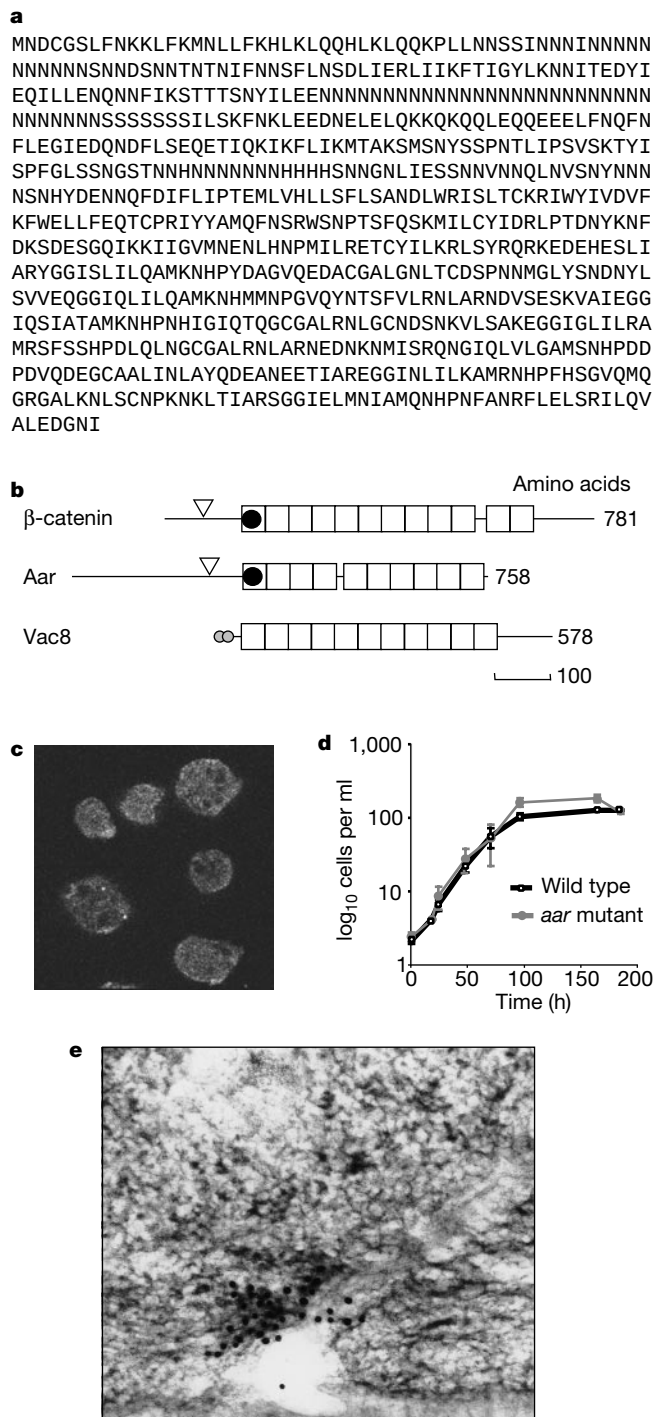


Figure 3 *aardvark* (*aar*) encodes a homologue of β -catenin. **a**, Sequence of Aardvark protein (Genbank Accession number AF305417). **b**, Scale diagrams of Aar, mouse β -catenin and Vac8. Open boxes indicate Arm repeats; a black circle indicates putative α -catenin binding¹³; triangle indicates a cluster of potential GSK-3 sites and light grey circles indicate the acylation sites of Vac8p. **c**, Anti-Aar staining of growing cells. Cells were fixed and immunostained with anti-Aar antiserum and visualized using a Texas Red conjugated secondary antiserum. **d**, Growth curves measured for wild-type cells and a mutant that lacks *aar*. **e**, A wild-type electron microscopy section incubated with anti-Aar antiserum and anti-rabbit IgG conjugated colloidal gold. Gold particles cluster at the junctions.

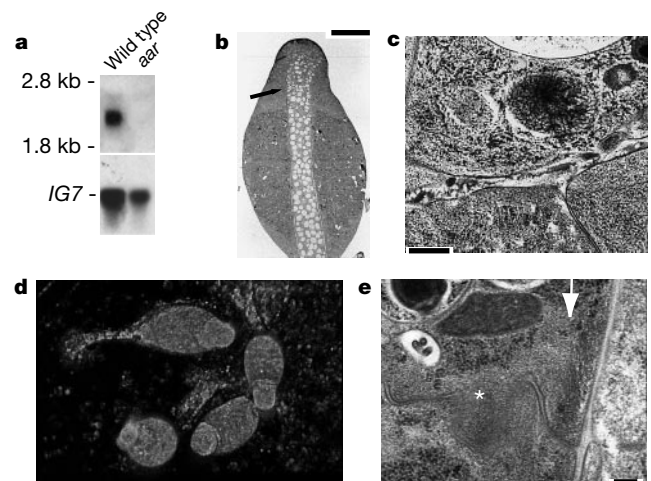


Figure 4 Analysis of a mutant that lacks the *aar* gene. **a**, Northern analysis shows the absence of *aar* mRNA in the *aar* mutant. *Ig7* is ubiquitously expressed and is shown as a loading control. **b**, Longitudinal section through a developing fruiting body (taken at 23 h), prepared as in Fig. 1a. Scale bar, 50 μ m. **c**, Electron micrograph of a transverse section through the region where cell junctions would be expected. Scale bar, 500 nm. **d**, A 3D-rendered image of ABPD-GFP expressing *aar* mutant cells shows that no actin ring is present in the absence of Aar protein. A series of whole fruiting bodies are shown after they have fallen to the substratum. **e**, Electron micrograph of a longitudinal section through an *aar* mutant overexpressing Aar. Asterisk indicates the position of an extended cell junction. White arrows marks position of an ectopic actin filament bundle. Scale bar, 500 nm.

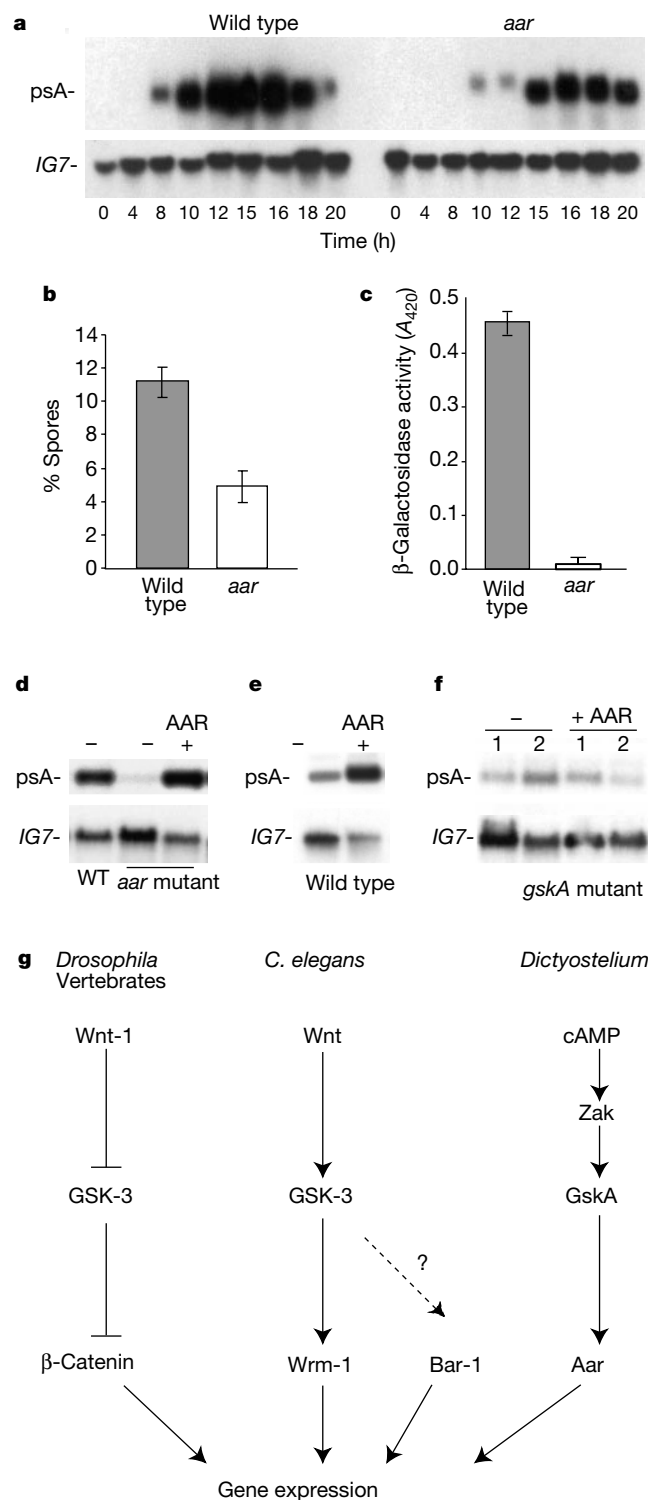


Figure 5 Aberrant cell differentiation in the *aar* mutant. **a**, Northern analysis showing expression of *psA* in wild type and the *aar* mutant. *IG7* expression indicates RNA loading. Time corresponds to time from initiation of development; multicellular development begins at 10 h and the fruiting body forms after 20 h. **b**, Spore-cell differentiation of isolated cells in monolayer culture. Graph shows mean percentage ($\pm$ s.e.m.) of spore cells formed after 24 h in the presence of 8-bromo-cAMP. **c**, **d**, Isolated cells were developed in suspension culture (5×10^6 cells ml^{-1}) for 8 h, and *psA* gene expression was induced by the addition of 5 mM cAMP for 12 h. Expression was detected by β -galactosidase activity in cells transformed with *psA::lacZ* (**c**) or by northern analysis of the endogenous *psA* gene (**d**). WT, wild-type cells; '+', expression of the *aar* cDNA from the *actin-15* promoter; '-', transformed with the plasmid vector. **e**, *psA* expression in

wild-type cells overexpressing the *aar* cDNA (+), developed as above. **f**, *gskA* mutant cells overexpressing the *aar* cDNA (+), developed as above. Two separate induction experiments (1 and 2) are shown. **g**, A comparison between β -catenin-mediated signal transduction pathways. In *Drosophila* and vertebrates, Wnt-1 stimulation counteracts GSK-3 β repression of β -catenin²⁰. In *C. elegans*, a Wnt signal acts positively on GSK-3 which then acts positively on the β -catenin homologue Wrm-1 (ref. 22). A second *C. elegans* β -catenin homologue is also involved in Wnt signalling but its relationship to GSK-3 is not known²³. In *Dictyostelium*, extracellular cAMP acts through the tyrosine kinase ZakA²⁷ to stimulate GskA activity¹⁹, which is required for Aar-mediated signalling. All pathways regulate gene expression.

evolutionary origin. First, the possession of adherens junctions is not a unique metazoan feature, and must have evolved before the emergence of the animals. Second, as Aar is also required for *psA* gene expression and full spore-cell differentiation, the signal transduction function of β -catenin may also have arisen before metazoan evolution and, as in animals, this signalling function requires GSK-3 activity. These observations are particularly interesting with respect to *C. elegans*. Here, separate β -catenin proteins undertake different functions, with the HMP-2 protein binding to cadherin; Bar-1 binding the transcription factor POP-1 and Wrm-1 acting through the kinase LIT-1 (ref. 23). Our results with *Dictyostelium* suggest a likely divergence in nematodes from a common protein possessing both cytoskeletal and signalling functions. In contrast, the only evidence to date suggests a positive genetic relationship between GSK-3 and β -catenin in *C. elegans*, resembling the situation in *Dictyostelium*, rather than that in *Drosophila* and vertebrates. Finally, a non-metazoan origin for both the structural and signal transduction functions of β -catenin, and the existence of related proteins in the plant *A. thaliana*, raises the possibility that the evolution of β -catenin may have been a prerequisite for all multicellular development. □

Methods

Cell culture and development

We grew cells at 22 °C in shaking culture in axenic medium and appropriate antibiotic selection. For development, logarithmically growing cells ($1\text{--}2 \times 10^6$ cells ml⁻¹) were washed in KK₂ (15.5 mM KH₂PO₄, 3.8 mM K₂HPO₄, pH 6.2) and plated on 0.45- μ m pore size nitrocellulose filters (Millipore) at a density of 5×10^7 cells per filter.

Construction of *aar* mutant strains

An *aar* mutant was generated by homologous recombination, using a plasmid made by replacing a 600-base-pair *Clal*/*EcoRI* fragment in the *aar* cDNA with a 1.4-kilobase (kb) *Clal*/*EcoRI* fragment containing a blasticidin-resistance gene expressed from an *actin-15* promoter. For *Aar* expression, we used a modified version of the pDXA-3C vector²⁴, in which the initiation codon in the polylinker had been removed (A.J.H., unpublished work). PCR was used to introduce a *KpnI* site into the 5' untranslated region and a *XhoI* site at the termination codon of the *aar* cDNA. This fragment was subcloned into the expression vector, so that it expresses under the control of the *actin-15* promoter, but translation initiates from the *aar* initiator codon. Clones were generated by electroporation of wild-type (AX2) or *aar* and *gskA* mutant cells²⁵.

Light microscopy

We used a purified 6 $\times$ His-tagged fusion protein containing the last five Arm repeats of Aar to immunize a New Zealand White male rabbit four times at 14-day intervals. The antiserum was used to stain growing *Dictyostelium* cells fixed in picric acid for 30 mins. Cells were incubated with anti-Aar for 3–4 h at room temperature, washed and incubated with Texas Red conjugated goat anti-rabbit antiserum (Vector Laboratories) for 1–2 h at room temperature before washing in PBS and water. Wild-type (AX2) and *aar* mutant cells were transformed with a plasmid expressing ABPD-GFP from the *actin-15* promoter⁹. Fluorescence and GFP localization was observed by confocal microscopy (Biorad MRC 600) on an inverted microscope (Nikon Diaphot). Z-series data was deconvolved and 3D rendered using Openlab 2.1 (Improvision).

Sample preparation and transmission electron microscopy

Developed wild-type (strain V12 M2) and mutant cells were fixed by ultrarapid plunge freezing in liquid propane cooled by liquid nitrogen. For ultrastructure, freeze substitution was carried out by placing the structures in 10 ml acetone/1% osmium tetroxide at -80 °C and samples were returned to room temperature over several days. Samples were embedded in Spurr's resin and ultrathin sections (70 nm) collected onto Formvar-coated nickel grids to be examined at 75 kV in a Hitachi HH-11E transmission electron microscope. For immunocytochemistry, freeze substitution was carried out in acetone alone and samples were embedded in Lowicryl HM20. Sections were blocked with TBS/BSA followed by primary antibody (either anti-actin monoclonal antibody, clone 4; ICN Biomedicals, Inc., Aurora, Ohio; or anti-Aar antiserum). Samples were washed with TBS and incubated with their respective IgG secondary antibodies conjugated to 10-nm colloidal gold particles (Amersham Life Sciences, Arlington Heights, Illinois)²⁶. All samples were post-stained with uranyl acetate and lead citrate.

Induction of prespore and spore cells

Cells were plated at low density in monolayer cultures as described previously¹⁸. Spore cells were induced by plating amoebae at 1.5×10^3 cells cm⁻² in spore medium supplemented with 15 mM 8-bromo-cAMP and incubating for 48 h at 22 °C. Spore cells were counted and expressed as a percentage of the total cell count. *psA* was induced by shaking cells at a

density of 5×10^6 cells ml⁻¹ for 6 h in KK₂, followed by 16 h with 5 mM cAMP. Induced cells were analysed by northern analysis or a *lacZ* assay as described¹⁸. Transformant controls contain the vector plasmid without the *aar* cDNA.

Received 8 August; accepted 18 September 2000.

- Stevenson, B. R. & Paul, D. L. The molecular constituents of intercellular junctions. *Curr. Opin. Cell Biol.* **1**, 884–891 (1989).
- Ozawa, M., Ringwald, M. & Kemler, R. Uvomorulin–catenin complex formation is regulated by a specific domain in the cytoplasmic region of the cell adhesion molecule. *Proc. Natl Acad. Sci. USA* **87**, 4246–4250 (1990).
- Ozawa, M., Baribault, H. & Kemler, R. The cytoplasmic domain of the cell adhesion molecule uvomorulin associates with three independent proteins structurally related in different species. *EMBO J.* **8**, 1711–1717 (1989).
- Nagafuchi, A. & Takeichi, M. Cell binding function of E-cadherin is regulated by the cytoplasmic domain. *EMBO J.* **7**, 3679–3684 (1988).
- Behrens, J. et al. Functional interaction of β -catenin with the transcription factor LEF-1. *Nature* **382**, 638–642 (1996).
- Molenaar, M. et al. XTcf-3 transcription factor mediates β -catenin-induced axis formation in *Xenopus* embryos. *Cell* **86**, 391–399 (1996).
- Huber, O. et al. Nuclear localization of β -catenin by interaction with transcription factor LEF-1. *Mech. Dev.* **59**, 3–10 (1996).
- Raper, K. B. *Dictyostelium discoideum*, a new species of slime mold from decaying forest leaves. *J. Agr. Res.* **50**, 135–147 (1935).
- Pang, K. M., Lee, E. & Knecht, D. A. Use of a fusion protein between GFP and an actin-binding domain to visualize transient filamentous-actin structures. *Curr. Biol.* **8**, 405–408 (1998).
- Nathke, I. S., Hinck, L., Swedlow, J. R., Papkoff, J. & Nelson, W. J. Defining interactions and distributions of cadherin and catenin complexes in polarized epithelial cells. *J. Cell Biol.* **125**, 1341–1352 (1994).
- Morio, T. et al. The *Dictyostelium* developmental cDNA project: generation and analysis of expressed sequence tags from the first-finger stage of development. *DNA Res.* **5**, 335–340 (1998).
- Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
- Aberle, H., Schwartz, H., Hoschuetzky, H. & Kemler, R. Single amino acid substitutions in proteins of the armadillo gene family abolish their binding to α -catenin. *J. Biol. Chem.* **271**, 1520–1526 (1996).
- Wang, Y. X., Catlett, N. L. & Weisman, L. S. Vac8p, a vacuolar protein with armadillo repeats, functions in both vacuole inheritance and protein targeting from the cytoplasm to vacuole. *J. Cell Biol.* **140**, 1063–1074 (1998).
- Fleckenstein, D., Rohde, M., Klionsky, D. J. & Rüdiger, M. Yel013p (Vac8p), an armadillo repeat protein related to plakoglobin and importin- α , is associated with the yeast vacuole membrane. *J. Cell Sci.* **111**, 3109–3118 (1998).
- Pan, X. & Goldfarb, D. S. YEB3/VAC8 encodes a myristylated armadillo protein of the *Saccharomyces cerevisiae* vacuolar membrane that functions in vacuole fusion and inheritance. *J. Cell Sci.* **111**, 2137–2147 (1998).
- Williams, J. G. et al. Origins of the prestalk-prespore pattern in *Dictyostelium* development. *Cell* **59**, 1157–1163 (1989).
- Harwood, A. J., Plyte, S. E., Woodgett, J., Strutt, H. & Kay, R. R. Glycogen synthase kinase 3 regulates cell fate in *Dictyostelium*. *Cell* **80**, 139–148 (1995).
- Plyte, S. E., O'Donovan, E., Woodgett, J. R. & Harwood, A. J. Glycogen synthase kinase-3 (GSK-3) is regulated during *Dictyostelium* development via the serpentine receptor cAR3. *Development* **126**, 325–333 (1999).
- Sokol, S. Y. Wnt signaling and dorso-ventral axis specification in vertebrates. *Curr. Opin. Genet. Dev.* **9**, 405–410 (1999).
- Rocheleau, C. E. et al. Wnt signaling and an APC-related gene specify endoderm in early *C. elegans* embryos. *Cell* **90**, 707–716 (1997).
- Schlesinger, A., Shelton, C. A., Maloof, J. N., Meneghini, M. & Bowerman, B. Wnt pathway components orient a mitotic spindle in the early *Caenorhabditis elegans* embryo without requiring gene transcription in the responding cell. *Genes Dev.* **13**, 2028–2038 (1999).
- Korswagen, H. C., Herman, M. A. & Clevers, H. C. Distinct β -catenins mediate adhesion and signalling functions in *C. elegans*. *Nature* **406**, 527–532 (2000).
- Manstein, D. J., Schuster, H. P., Morandini, P. & Hunt, D. M. Cloning vectors for the production of proteins in *Dictyostelium discoideum*. *Gene* **162**, 129–134 (1995).
- Howard, P. K., Ahern, K. G. & Firtel, R. A. Establishment of a transient expression system for *Dictyostelium discoideum*. *Nucleic Acids Res.* **16**, 2613–2623 (1988).
- Strafstrom, J. & Steahelin, L. Antibody localisation of extensin in cell walls of carrot storage roots. *Planta* **174**, 321–332 (1988).
- Kim, L., Liu, J. & Kimmel, A. R. The novel tyrosine kinase ZAK1 activates GSK3 to direct cell fate specification. *Cell* **99**, 399–408 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>).

Acknowledgements

We thank A. Staehelin, who suggested that we examine the stalk pinch region for an actin ring. We would also like to thank R. Williams and E. Dalton for their help with making DNA constructs; and D. Knight, V. Braga, R. Kay and J. Hyams for their advice and comments. J.C.C. was supported by a MRC/GlaxoWellcome Studentship. J.P.R. is a Wellcome Prize Student and A.J.H. is a Wellcome Trust Senior Fellow. This work was partly funded by a NATO collaborative research grant, the research enhancement fund of the College of Arts and Sciences, Texas Tech University, and departmental and university funds to the Department of Biological Sciences Electron Microscopy Laboratory at Texas Tech University.

Correspondence and requests for materials should be addressed to A.J.H. (e-mail: a.harwood@ucl.ac.uk).

IRSp53 is an essential intermediate between Rac and WAVE in the regulation of membrane ruffling

Hiroaki Miki, Hideki Yamaguchi, Shiro Suetsugu & Tadaomi Takenawa

Department of Biochemistry, Institute of Medical Science, University of Tokyo, and CREST, Japan Science and Technology Corporation, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Neural Wiskott-Aldrich syndrome protein (N-WASP)¹ functions in several intracellular events including filopodium formation², vesicle transport³ and movement of *Shigella flexneri*^{4,5} and vaccinia virus⁶, by stimulating rapid actin polymerization through the Arp2/3 complex. N-WASP is regulated by the direct binding of Cdc42 (refs 7, 8), which exposes the domain in N-WASP that activates the Arp2/3 complex^{2,9}. A WASP-related protein, WAVE/Scar, functions in Rac-induced membrane ruffling^{10,11}; however, Rac does not bind directly to WAVE¹⁰, raising the question of how WAVE is regulated by Rac. Here we demonstrate that IRSp53, a substrate for insulin receptor¹² with unknown function, is the

'missing link' between Rac and WAVE. Activated Rac binds to the amino terminus of IRSp53, and carboxy-terminal Src-homology-3 domain of IRSp53 binds to WAVE to form a trimolecular complex. From studies of ectopic expression, we found that IRSp53 is essential for Rac to induce membrane ruffling, probably because it recruits WAVE, which stimulates actin polymerization mediated by the Arp2/3 complex.

We have previously found and reported three WAVEs (WAVE1, -2 and -3)¹³. All WAVEs contain an N-terminal SHD/WH2 (Scar-/WAVE-homology domain)¹⁴, a central proline-rich region and a C-terminal VCA region (for verprolin-homology, cofilin-homology and acidic-residue-rich regions) that stimulates actin nucleation through Arp2/3 complex¹⁵. As proline-rich regions are known to be involved in protein-protein interactions, we carried out a two-hybrid screening using the proline-rich region of WAVE1 as 'bait', which yielded several positive clones. DNA sequencing analysis revealed that they encode three identified proteins (IRSp53, nArgBP2 and profilin) and two unidentified ones. As IRSp53 gave the strongest positive signal on β -galactosidase assay (data not shown), we analysed it further.

We first confirmed that IRSp53 and WAVE were co-immunoprecipitated by antibodies for IRSp53 and WAVE (see Supplementary Information for the specificity of the anti-IRSp53 antibody), indicating their *in vivo* complex formation (Fig. 1a). N-WASP was

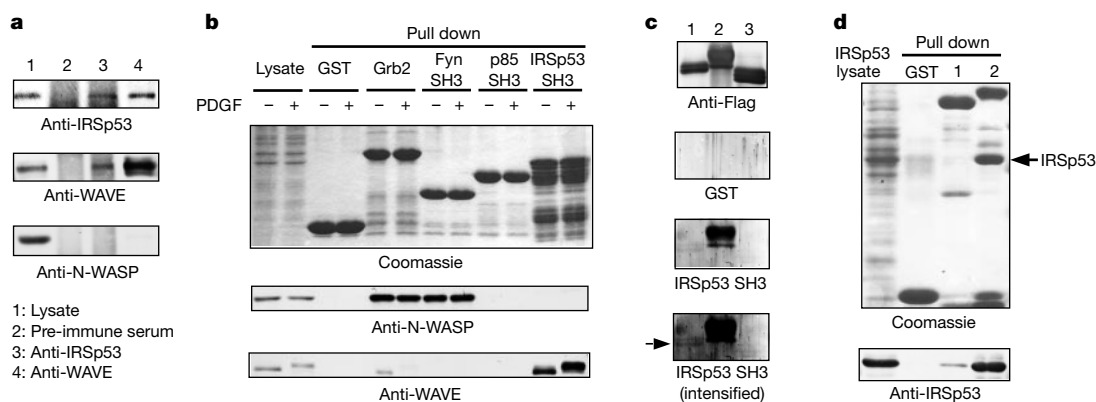


Figure 1 The specific and direct interaction of IRSp53 and WAVE2. **a**, NIH3T3 cell lysates were immunoprecipitated with anti-IRSp53 or anti-WAVE antibodies. The precipitates were analysed by immunoblotting. **b**, Various GST-fusion SH3 domain-containing proteins immobilized on beads were mixed with Swiss3T3 cell lysates treated with or without PDGF (WAVE mobility-shift occurs by PDGF-treatment²³). The bound proteins

were analysed by immunoblotting. **c**, The expressed Flag-WAVE1, -2 and -3 were immunoprecipitated with anti-Flag antibody and analysed by far-western blot with GST-IRSp53 SH3 (100 nM). Detection was by an anti-GST antibody. **d**, Full-length IRSp53 (about 2 μ M) was subjected to a pull-down assay with GST-fusion proteins of full-length WAVE1 and -2.

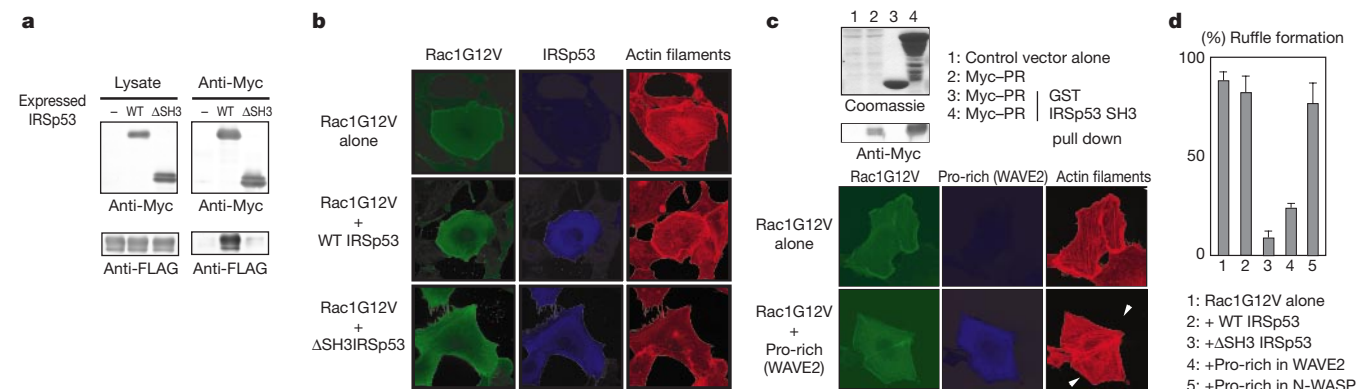


Figure 2 Importance of IRSp53/WAVE2 in Rac-induced membrane ruffling. **a**, Myc-IRSp53 (WT and Δ SH3) and Flag-WAVE2 were co-expressed in COS7 cells, and then immunoprecipitated with anti-Myc antibody. The precipitates were analysed by immunoblotting. **b**, Flag-Rac1G12V was expressed in NIH3T3 cells with Myc-IRSp53 (WT and Δ SH3). The cells were then subjected to immunofluorescence microscopic analyses by staining with anti-Flag antibody (green), anti-Myc antibody (blue) and

phalloidin (red). **c**, Top, Myc-PR was expressed and subjected to a pull-down assay with GST-IRSp53 SH3. Bottom, Myc-PR was expressed with Flag-Rac1G12V in NIH3T3 cells and then stained as described in **b**. **d**, Cells with membrane ruffles over half of the cell periphery were counted as ruffling positive. Results shown are averages from two independent experiments (50 cells were counted in each experiment). Error bars represent the difference in each experiment from the average.

hardly precipitated. We then examined whether the Src-homology (SH)-3 domain, located at the C terminus of IRSp53, associates with WAVE by a pull-down assay. The SH3 domain in IRSp53 strongly associated with WAVE, but N-WASP was not precipitated (Fig. 1b). This is comparable to Grb2/Ash and Fyn, which associate strongly with N-WASP, but less significantly with WAVE (Fig. 1b).

The anti-WAVE antibody that we used recognizes all three WAVES (data not shown); therefore, we next investigated their interaction using ectopically expressed Flag-WAVE1, -2 and -3. We immunoprecipitated Flag-WAVE with anti-Flag antibody and then analysed the precipitates by far-western blot with the SH3 protein (Fig. 1c). Strong and direct binding to WAVE2 was observed, but WAVE1 gave a very weak positive signal that can be seen in the intensified image (arrowhead). To confirm the binding specificity under a more physiological condition, we immobilized full-length WAVE1 and -2 and examined their binding ability to full-length IRSp53 (Fig. 1d). Again, WAVE2 pulled down a significantly larger amount of IRSp53, which is clearly seen even in the Coomassie stained image. These data suggest that WAVE2 is the physiological target for IRSp53. We further pinpointed the IRSp53-binding region in WAVE2 (see Fig. 2 in Supplementary information). WAVE2 has been reported to be expressed ubiquitously, whereas WAVE1 and 3 are found mostly in brain¹³. We determined what type of WAVE(s) exists in fibroblasts by pull-down assay using the IRSp53 SH3 domain. We found that more than 95% of WAVES were precipitated, indicating that WAVE2 is the principal WAVE in fibroblasts (data not shown).

To examine the functional significance of the binding between IRSp53 and WAVE2, we prepared a mutant IRSp53 lacking the SH3

domain (Δ SH3). This mutation markedly reduced the ability of IRSp53 to bind WAVE2 (Fig. 2a). We then co-expressed IRSp53 (wild type and Δ SH3) and activated Rac (Rac1G12V). As shown previously, the expression of activated Rac induced membrane ruffling in fibroblasts¹¹, and co-expression of wild-type IRSp53 had no effect (Fig. 2b). However, co-expression of Δ SH3 IRSp53 clearly suppressed the Rac-induced membrane ruffling. This inhibitory effect was specific to Rac, as the co-expression of Δ SH3 IRSp53 had no effect on either Cdc42-induced actin microspike formation or Rho-induced stress fibre formation (data not shown).

If IRSp53 and WAVE2 binding is critical for Rac-induced membrane ruffling, then inhibition of this binding will show a dominant-negative effect. As a binding competitor, we used a partial fragment of the proline-rich region in WAVE2 (PR; amino acids 273–402), which strongly associated with the IRSp53 SH3 domain (Fig. 2c, top). The co-expression of PR with activated Rac significantly suppressed membrane ruffling, and actually induced strong stress fibre formation (Fig. 2c, bottom). These stress fibres are probably induced by the activation of endogenous Rho by activated Rac as previously reported¹¹. The co-expression of Δ SH3 IRSp53 suppressed ruffling but did not show any enhancement of stress fibre formation, which may reflect that they inhibit the Rac signalling at different points (Fig. 2b). We also confirmed that a similar proline-rich fragment of N-WASP did not suppress Rac-induced membrane ruffling (Fig. 2d). Clearly, Δ SH3 IRSp53, an IRSp53 mutant that does not bind WAVE2, and the PR region in WAVE2, which is a binding interface to IRSp53, significantly inhibit Rac-induced membrane ruffling, indicating that the IRSp53–WAVE2

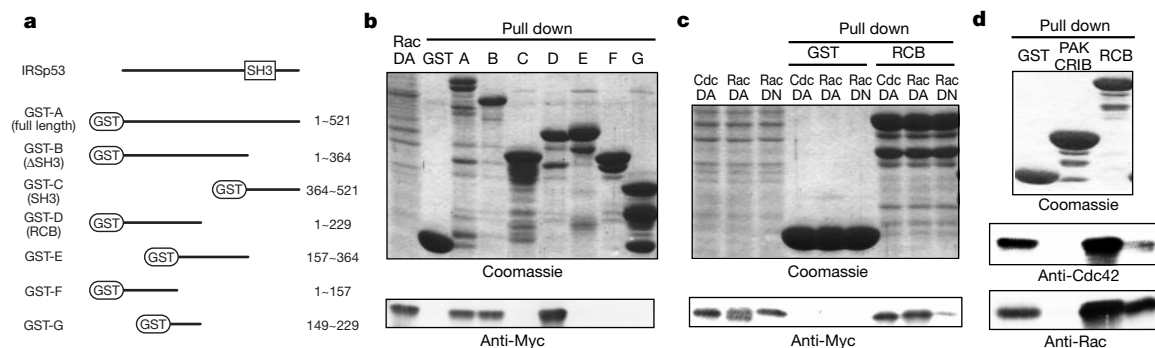


Figure 3 Direct association of activated Rac with the N-terminal region of IRSp53 (RCB). **a**, Structures of IRSp53 GST-fusion proteins. Amino-acid numbers are indicated. **b**, Myc–Rac1G12V (RacDA) was expressed and subjected to a pull-down assay with indicated GST-fusion proteins. **c**, The expressed Myc–Cdc42G12V (CdcDA),

–Rac1G12V (RacDA) and –Rac1T17N (RacDN) were subjected to a pull-down assay with GST–RCB. **d**, Purified Cdc42 and Rac proteins (1 μ M) pre-loaded with GTP were subjected to a pull-down assay with GST–RCB. Bound Cdc42 and Rac were detected with anti-Cdc42 and anti-Rac antibodies, respectively. GST–PAK CRIB is used as a positive control.

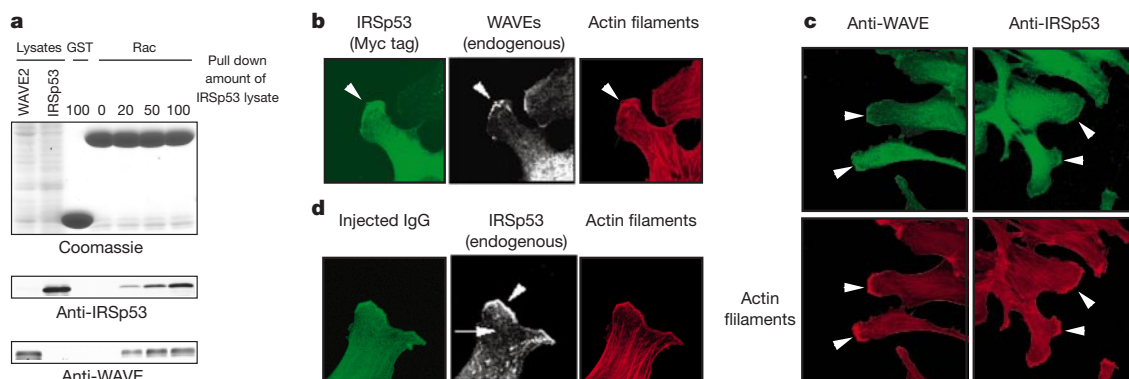


Figure 4 The IRSp53-mediated association of Rac and WAVE2. **a**, GST–Rac (GTP-loaded) immobilized on beads was incubated with cell lysates expressing WAVE2 (400 μ l) in the presence of indicated amounts of lysates expressing IRSp53. **b**, Myc–IRSp53 was expressed in NIH3T3 cells, and then cells were made polarized with wounds. After 8 h, they were stained with anti-Myc antibody (green), anti-WAVEs antibody (blue) and phalloidin (red). **c**, Polarized NIH3T3 cells were stained with phalloidin (red) and the

anti-IRSp53 antibody or the anti-WAVEs antibody (green). **d**, Polarized NIH3T3 cells were injected with mouse IgG and cultured for 2 h. Cells were then stained with anti-mouse IgG antibody (green), anti-IRSp53 antibody (blue) and phalloidin (red). Arrowhead, accumulation of endogenous IRSp53 at the edge of ruffles; arrow, cytoplasmic area where IRSp53 signal is weak.

complex has a crucial role in this process.

We thought that IRSp53 directly links Rac to WAVE2 and therefore checked for possible interactions between IRSp53 and Rac. Glutathione S-transferase fusion proteins of IRSp53 (see Fig. 3a) did pull down activated Rac protein expressed in COS7 cells (Fig. 3b). Δ SH3 (GST-B) bound to Rac, while SH3 domain itself (GST-C) did not, indicating that SH3 domain is not involved in this binding. The smallest binding fragment was GST-D, which contains the N-terminal 229 residues ('RCB', for Rac-binding). Further deleted fragments from either the N or C terminus (GST-E, -F and -G) lost the binding ability, although it is possible that these proteins did not fold correctly. We carried out an extensive database search for a homologous sequence to this RCB region, but did not find significant homology to any sequence.

To confirm the binding specificity, we performed a similar pull-down experiment using activated Cdc42 (Cdc42G12V) and inactivated Rac (Rac1T17N) protein against RCB. Inactivated Rac was hardly precipitated, but the amount of activated Cdc42 precipitated was comparable to that of activated Rac in this assay (Fig. 3c). To determine whether the binding was direct, we performed an assay using purified Cdc42 and Rac, but not crude cell lysates. After GTP loading, the proteins were mixed with immobilized RCB. As shown in Fig. 3d, GTP-loaded Rac associated directly with RCB at similar efficiency to the CRIB motif of human PAK1 (amino acids 31–100), which was used as a positive control¹⁶. In the case of Cdc42, however, the binding efficiency was much weaker than that of the PAK CRIB

motif. These results confirm that IRSp53 is a direct target of Rac. IRSp53 may also bind to Cdc42 in cells, but Cdc42 can induce rapid actin polymerization through the well-characterized N-WASP pathway⁹.

We next investigated whether IRSp53 can link Rac to WAVE2. When Rac (GTP-loaded) immobilized on beads was mixed with WAVE2, there was no precipitation of WAVE2 (Fig. 4a), indicating that there is no direct binding between them. When Rac-beads were mixed with WAVE2 in the presence of various amounts of IRSp53, however, WAVE2 was co-precipitated with the beads. Further, the precipitated amount of WAVE2 correlated well with that of IRSp53, clearly showing that IRSp53 physically links Rac to WAVE2.

If IRSp53 is a linker of WAVE2, these proteins should co-localize in cells, especially in areas where membrane ruffling occurs. To confirm this possibility, we examined the localization of expressed IRSp53 (detected by monoclonal anti-Myc antibody) and endogenous WAVEs (mostly WAVE2) in fibroblasts. A clear co-localization was observed at ruffles (Fig. 4b). We then compared the localization pattern of endogenous IRSp53 and WAVEs. Both proteins show a dot-like pattern in the cytoplasm and an accumulation at ruffles (Fig. 4c). Both antibodies are rabbit polyclonal ones, and thus it was impossible to compare their localization in a single cell. In general, ruffling areas become thick, which suggests the possibility that the 'localization' of IRSp53 and WAVEs at ruffles may only reflect the cytoplasmic 'thickness'. To check this possibility, we microinjected mouse immunoglobulin- γ (IgG) as a cytoplasmic marker and then observed the localization of endogenous IRSp53 (Fig. 4d). As expected, some accumulation of IgG was observed at ruffling areas; however, the IRSp53 signal is very strong at the edges of ruffles (arrowhead) but very weak in other areas (arrow), showing a striking contrast to IgG localization. Staining of endogenous WAVEs also yielded a similar result (data not shown), supporting the notion that IRSp53 and WAVEs do not passively accumulate at ruffles after they are formed.

Finally, we examined the possible effects of IRSp53 binding on WAVE2-induced Arp2/3 complex activation in a pyrene actin assay. The VCA region in WAVE2 stimulated Arp2/3-complex-mediated actin polymerization (Fig. 5a), as did those in WASP¹⁷, N-WASP⁹, and WAVE1 (ref. 15). The full-length protein of WAVE2 also showed the activity, and there seemed no significant difference in the activating intensity as compared with the VCA region alone. We pre-incubated IRSp53 protein with full-length WAVE2 before purification, and recovered the two in a complex form (about 50% of WAVE2 is estimated to be complexed with IRSp53). When this IRSp53–WAVE2 complex was subjected to assay, it showed a much stronger effect than WAVE2 alone (Fig. 5a). In addition, this stimulatory effect was observed to be dose-dependent when purified WAVE2 and IRSp53 were mixed 5 min before the assay (Fig. 5b). Addition of IRSp53 to WAVE1 or the VCA region of WAVE2, both of which do not bind IRSp53 significantly, had neither a positive nor negative effect (Fig. 5c). Therefore, the complex formation of WAVE2 with IRSp53 has a significant stimulatory effect on WAVE2-induced actin polymerization through Arp2/3 complex. We also examined the possible effect of GTP-loaded Rac on IRSp53–WAVE2 complex, but did not find any positive/negative effect on the actin polymerization kinetics (data not shown).

We have shown that IRSp53 is the critical linker between Rac and WAVE2, an Arp2/3 complex activator. The activated Arp2/3 complex stimulates *de novo* actin nucleation, leading to rapid actin polymerization with branches^{18–20}, which are typically observed in ruffles^{21,22}. IRSp53 also activates WAVE2 in an Arp2/3-complex-mediated actin polymerization assay. The level of activation is comparable to that of fully activated N-WASP (data not shown), which indicates that Rac can recruit a strong activator for the Arp2/3 complex. We do not know, however, whether this activation has a physiological relevance—an issue that should be investigated in future studies.

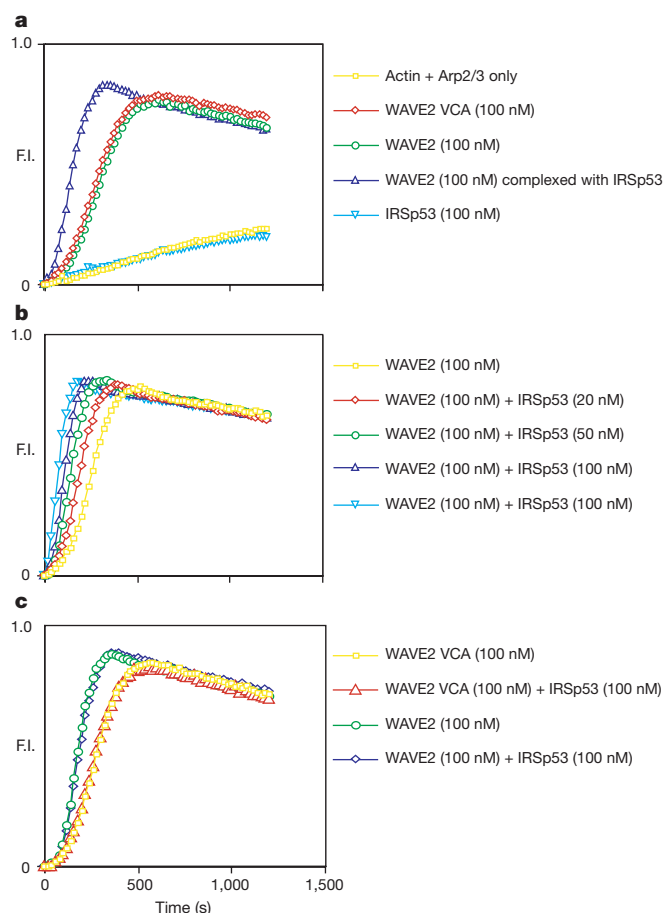


Figure 5 Stimulation of Arp2/3 complex activation by WAVE2 in the presence of IRSp53. **a**, Actin polymerization kinetics were examined using pyrene-labelled actin. Pyrene-labelled monomeric actin (0.2 μ M), free monomeric actin (2 μ M), Arp2/3 complex (60 nM) purified from bovine brain, and the indicated proteins (WAVE2 and WAVE2 VCA are GST-fusion forms) were mixed in assay buffer, and the fluorescence intensity (F.I.) was monitored. **b**, 20 nM, 50 nM, 100 nM, 200 nM, and 400 nM IRSp53 with 100 nM WAVE2 were tested. **c**, Effects of IRSp53 on WAVE1 and WAVE2 VCA were tested.

Methods

Two-hybrid screening

We used the Matchmaker Two-Hybrid system (Clontech). DNA coding the proline-rich region of WAVE1 was amplified by PCR and subcloned in pGBT9 plasmid vector. This recombinant plasmid was transformed into Y190 yeast, which was then used as a host cell for screening. A human brain cDNA library (Clontech) was introduced into the transformed yeast and selected.

Recombinant proteins

We expressed various partial fragments of IRSp53 as GST–fusion forms in *Escherichia coli* using pGEX plasmids (Pharmacia). GST–fusion proteins of Grb2/Ash, Fyn, and p85 (phosphatidylinositol 3-kinase) were prepared as described²³. Cdc42 and Rac were expressed as GST–fusion forms in Sf9 cells using recombinant baculoviruses, which were produced using the BAC-TO-BAC system (Gibco BRL). After purification with glutathione-sepharose beads, GST was cleaved off by thrombin treatment. Full-length WAVE1, WAVE2 and IRSp53 were also expressed as either GST–fusion forms or non-tagged forms in Sf9 cells by recombinant baculoviruses.

Antibodies

Rabbit polyclonal anti-WAVE antibody was prepared as described¹⁰. Anti-IRSp53 antibody was prepared in a rabbit immunized with the N-terminal 157-residue proteins expressed in *E. coli*. The antibody was affinity purified. We used commercially available antibodies for Myc-tag (polyclonal and monoclonal, both from Santa Cruz), Flag-tag (monoclonal from Sigma), Cdc42 (polyclonal from Santa Cruz), Rac (monoclonal from Transduction Laboratory) and GST (polyclonal from Santa Cruz). Control mouse IgG was purchased from Sigma.

Pull-down assay

GST–fusion proteins (10–50 µg) were first immobilized on 20 µl of glutathione-sepharose beads and then mixed with 400 µl protein samples such as cell lysates and purified proteins. After 2 h, the beads were washed with lysis buffer five times and 20 µl of SDS sample buffer was added. Samples (5 µl) were separated by SDS–PAGE, followed by immunoblotting and Coomassie staining. Cell lysates of Swiss3T3, COS7 and NIH3T3 cells were obtained by lysing half-confluent cells in a dish (diameter, 150 mm) with 1 ml lysis buffer. In the case of Sf9 cells infected with baculoviruses, cells were collected by centrifugation and then lysed in 1/20 volume of lysis buffer.

Ectopic expression in mammalian cells

NIH3T3 cells were transfected by the Ca²⁺-phosphate method as described²³. COS7 cells were transfected by electroporation as described². In all expression analyses, pEF-BOS plasmid vectors² were used.

Pyrene actin assay

Actin was purified from rabbit muscle. We carried out pyrene labelling as described⁹. Arp2/3 complex was purified by affinity chromatography as described⁹. Arp2/3 complex and various proteins except actin were mixed in assay buffer (2 mM Tris-HCl (pH 8.0), 2 mM MgCl₂, 100 mM KCl, 0.2 mM CaCl₂, 0.2 mM ATP and 0.5 mM dithiothreitol). After incubation for 5 min at room temperature, actin (0.2 µM labelled actin in 2 µM non-labelled actin) was added and then subjected to fluorometry. Excitation and emission wavelengths were 365 nm and 407 nm, respectively.

Received 20 July; accepted 1 September 2000.

1. Miki, H., Miura, K. & Takenawa, T. N-WASP, a novel actin-depolymerizing protein, regulates the cortical cytoskeletal rearrangement in a PIP2-dependent manner downstream of tyrosine kinases. *EMBO J.* **15**, 5326–5335 (1996).
2. Miki, H. *et al.* Induction of filopodium formation by a WASP-related actin-depolymerizing protein N-WASP. *Nature* **391**, 93–96 (1998).
3. Rozelle, A. L. *et al.* Phosphatidylinositol 4,5-bisphosphate induces actin-based movement of raft-enriched vesicles through WASP-Arp2/3. *Curr. Biol.* **10**, 311–320 (2000).
4. Suzuki, T. *et al.* Neural Wiskott–Aldrich syndrome protein is implicated in the actin-based motility of *Shigella flexneri*. *EMBO J.* **17**, 2767–2776 (1998).
5. Egile, C. *et al.* Activation of the CDC42 effector N-WASP by the *Shigella flexneri* IcsA protein promotes actin nucleation by Arp2/3 complex and bacterial actin-based motility. *J. Cell Biol.* **146**, 1319–1332 (1999).
6. Frischknecht, F. *et al.* Actin-based motility of vaccinia virus mimics receptor tyrosine kinase signalling. *Nature* **401**, 926–929 (1999).
7. Nobes, C. D. & Hall, A. Rho, rac, and cdc42 GTPases regulate the assembly of multimolecular focal complexes associated with actin stress fibers, lamellipodia, and filopodia. *Cell* **81**, 53–62 (1995).
8. Kozma, R. *et al.* The Ras-related protein Cdc42Hs and bradykinin promote formation of peripheral actin microspikes and filopodia in Swiss 3T3 fibroblasts. *Mol. Cell. Biol.* **15**, 1942–1952 (1995).
9. Rohatgi, R. *et al.* The interaction between N-WASP and the Arp2/3 complex links Cdc42-dependent signals to actin assembly. *Cell* **97**, 221–231 (1999).
10. Miki, H., Suetsugu, S. & Takenawa, T. WAVE, a novel WASP-family protein involved in actin reorganization induced by Rac. *EMBO J.* **17**, 6932–6941 (1998).
11. Ridley, A. J. *et al.* The small GTP-binding protein rac regulates growth factor-induced membrane ruffling. *Cell* **70**, 401–410 (1992).
12. Yeh, T. C. *et al.* Characterization and cloning of a 58/53-kDa substrate of the insulin receptor tyrosine kinase. *J. Biol. Chem.* **271**, 2921–2928 (1996).
13. Suetsugu, S., Miki, H. & Takenawa, T. Identification of two human WAVE/SCAR homologues as general actin regulatory molecules which associate with the Arp2/3 complex. *Biochem. Biophys. Res. Commun.* **260**, 296–302 (1999).

14. Bear, J. E., Rawls, J. F. & Saxe, C. L. III SCAR, a WASP-related protein, isolated as a suppressor of receptor defects in late *Dictyostelium* development. *J. Cell Biol.* **142**, 1325–1335 (1998).
15. Machesky, L. M. *et al.* Scar, a WASP-related protein, activates nucleation of actin filaments by the Arp2/3 complex. *Proc. Natl Acad. Sci. USA* **96**, 3739–3744 (1999).
16. Burbelo, P. D., Drechsel, D. & Hall, A. A conserved binding motif defines numerous candidate target proteins for both Cdc42 and Rac GTPases. *J. Biol. Chem.* **270**, 29071–29074 (1995).
17. Yasar, D. *et al.* The Wiskott–Aldrich syndrome protein directs actin-based motility by stimulating actin nucleation with the Arp2/3 complex. *Curr. Biol.* **9**, 555–558 (1999).
18. Mullins, R. D., Heuser, J. A. & Pollard, T. D. The interaction of Arp2/3 complex with actin: nucleation, high affinity pointed end capping, and formation of branching networks of filaments. *Proc. Natl Acad. Sci. USA* **95**, 6181–6186 (1998).
19. Blanchoin, L. *et al.* Direct observation of dendritic actin filament networks nucleated by Arp2/3 complex and WASP/Scar proteins. *Nature* **404**, 1007–1011 (2000).
20. Pantaloni, D. *et al.* The Arp2/3 complex branches filament barbed ends functional antagonism with capping proteins. *Nature Cell Biol.* **2**, 385–391 (2000).
21. Bailly, M. *et al.* Relationship between Arp2/3 complex and the barbed ends of actin filaments at the leading edge of carcinoma cells after epidermal growth factor stimulation. *J. Cell Biol.* **145**, 331–345 (1999).
22. Svitkina, T. M. & Borisy, G. G. Arp2/3 complex and actin depolymerizing factor/cofilin in dendritic organization and treadmilling of actin filament array in lamellipodia. *J. Cell Biol.* **145**, 1009–1026 (1999).
23. Miki, H. *et al.* Phosphorylation of WAVE downstream of mitogen-activated protein kinase signaling. *J. Biol. Chem.* **274**, 27605–27609 (1999).

Acknowledgements

We thank K. Miki for technical assistance in yeast two-hybrid screening and characterization of IRSp53, and N. Sasaki for WAVE1 proteins and recombinant baculoviruses. We are also grateful to T. Suzuki and C. Sasaki for the gift of the CRIB motif fragment of PAK. This study was supported in part by a Grant-in-Aid for Cancer Research from the Ministry of Education, Science, and Culture of Japan.

Correspondence and requests for materials should be addressed to T.T. (e-mail: takenawa@ims.u-tokyo.ac.jp).

InsP₄ facilitates store-operated calcium influx by inhibition of InsP₃ 5-phosphatase

Meredith C. Hermosura*, Hiroshi Takeuchi†, Andrea Fleig*, Andrew M. Riley‡, Barry V. L. Potter‡, Masato Hirata‡ & Reinhold Penner*

* Laboratory of Cell and Molecular Signaling, Center for Biomedical Research at The Queen's Medical Center and John A. Burns School of Medicine at The University of Hawaii, Honolulu, Hawaii 96813, USA

† Laboratory of Molecular and Cellular Biochemistry, Graduate School of Dental Sciences, Kyushu University and Kyushu University Station for Collaborative Research, Fukuoka 812-8582, Japan

‡ Wolfson Laboratory of Medicinal Chemistry, Department of Pharmacy & Pharmacology, University of Bath, Claverton Down, Bath BA2 7AY, UK

Receptor-mediated generation of inositol 1,4,5-trisphosphate (InsP₃) initiates Ca²⁺ release from intracellular stores and the subsequent activation of store-operated calcium influx¹. InsP₃ is metabolized within seconds by 5-phosphatase and 3-kinase², yielding Ins(1,4)P₂ and inositol 1,3,4,5-tetrakisphosphate (InsP₄), respectively. Some studies have suggested that InsP₄ controls Ca²⁺ influx in combination with InsP₃ (refs 3 and 4), but another study did not find the same result⁵. Some of the apparent conflicts between these previous studies have been resolved⁶; however, the physiological function of InsP₄ remains elusive^{7,8}. Here we have investigated the function of InsP₄ in Ca²⁺ influx in the mast cell line RBL-2H3, and we show that InsP₄ inhibits InsP₃ metabolism through InsP₃ 5-phosphatase, thereby facilitating the activation of the store-operated Ca²⁺ current I_{CRAC} (ref. 9). Physiologically, this mechanism opens a discriminatory time window for coincidence detection that enables selective facilitation of Ca²⁺ influx by appropriately timed low-level receptor stimulation. At higher

concentrations, InsP_4 acts as an inhibitor of InsP_3 receptors, enabling InsP_4 to act as a potent bi-modal regulator of cellular sensitivity to InsP_3 , which provides both facilitatory and inhibitory feedback on Ca^{2+} signalling.

Evidence suggests that a functionally distinct store is involved in activating I_{CRAC} and that depletion of this store requires fairly high InsP_3 levels^{10,11}. This lower InsP_3 sensitivity is probably due to InsP_3 metabolism through 5-phosphatase, which results in a nonlinear relationship between InsP_3 concentration and I_{CRAC} activation^{10,12}. As early work on 5-phosphatase identified InsP_4 as an inhibitor of this enzyme *in vitro*², we performed whole-cell patch-clamp experiments in which we perfused InsP_4 and other potential 5-phosphatase inhibitors in combination with InsP_3 . We found that inhibition of 5-phosphatase-dependent InsP_3 metabolism facilitates activation of I_{CRAC} .

Perfusion of 20 μM InsP_4 into RBL-2H3-M1 cells, a transfected cell line that stably expresses muscarinic M1 receptors¹³, produces no measurable Ca^{2+} release nor does it activate Ca^{2+} influx ($n = 12$; Fig. 1a). By contrast, carbachol-stimulated production of InsP_3 causes a sharp Ca^{2+} release transient followed by activation of store-operated Ca^{2+} entry. This carbachol-stimulated Ca^{2+} influx accounts for the Ca^{2+} increases during hyperpolarizing shifts in membrane potential. Similarly, under experimental conditions that favour electrophysiological detection of the store-operated Ca^{2+} current I_{CRAC} (ref. 9), InsP_4 (20 μM , $n = 12$) failed to activate

I_{CRAC} , whereas the same concentration of InsP_3 readily did so ($n = 36$; Fig. 1b).

We next assessed the effects of InsP_4 on I_{CRAC} in combination with defined InsP_3 concentrations. Activation of I_{CRAC} occurs over a concentration range of 2 to 5 μM (Fig. 1d; see also ref. 10); however, in the presence of InsP_4 the threshold for InsP_3 -mediated activation of I_{CRAC} is lowered considerably, enabling cells to develop the current even at subthreshold InsP_3 levels (1 μM and less). InsP_3 at concentrations of 2–5 μM activates I_{CRAC} with near-maximal amplitude, but with a delay and slow-activation time constant (Fig. 1d). At these InsP_3 concentrations, InsP_4 facilitates activation of I_{CRAC} by reducing the delay and by speeding-up activation of the current. Facilitation by InsP_4 , however, is no longer evident at high, saturating InsP_3 concentrations (above 5 μM ; Figs 1d and 2a). As InsP_4 is produced in numerous cell types, we tested whether the facilitation of I_{CRAC} also occurs in Jurkat T lymphocytes, another immune cell line with store-operated currents that are well characterized¹⁴. InsP_4 also facilitates I_{CRAC} activation induced by 1 μM InsP_3 in T cells (Fig. 1c), indicating that this mechanism may be of general importance.

The effects of InsP_4 on InsP_3 -mediated I_{CRAC} activation are shown in Fig. 2. First, InsP_4 lowers the threshold at which InsP_3 activates I_{CRAC} , which results in a significant left shift ($P < 0.05$) of the InsP_3 dose-response curve (Fig. 2a). Second, InsP_4 reduces delays in activation of the current, particularly at threshold levels (2–3 μM)

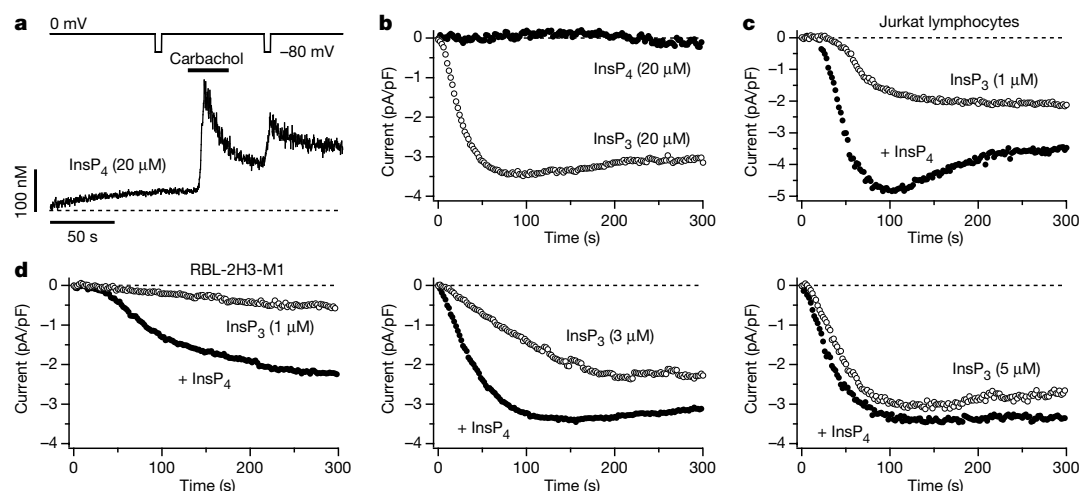


Figure 1 InsP_4 facilitates InsP_3 -dependent activation of I_{CRAC} . **a**, Representative $[\text{Ca}^{2+}]_i$ measurement in a patch-clamped RBL-2H3-M1 cell perfused with 20 μM InsP_4 . Identical hyperpolarizations to -80 mV (top trace) were induced before and after carbachol (100 μM) challenge to increase the driving force for Ca^{2+} and probe Ca^{2+} entry. **b**, Average inward currents of RBL-2H3-M1 cells at -80 mV in the presence of 20 μM InsP_4 ($n = 12$) and 20 μM InsP_3 ($n = 36$). **c**, Average inward currents at -80 mV of Jurkat

lymphocytes perfused with 1 μM InsP_3 ($n = 6$) or in combination with 20 μM InsP_4 ($n = 6$). **d**, Average inward currents of RBL-2H3-M1 cells at -80 mV in response to a given InsP_3 concentration or in combination with 20 μM InsP_4 . Experimental conditions in **b–d** are optimized for electrophysiological detection of I_{CRAC} (10 mM external Ca^{2+} and $[\text{Ca}^{2+}]_i$ buffered to 100 nM by 10 mM BAPTA + 4.3 mM CaCl_2).

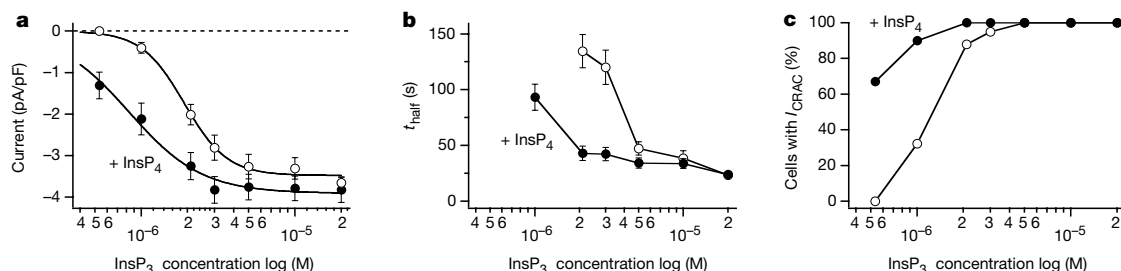


Figure 2 Quantitative analysis of InsP_4 -induced facilitation of I_{CRAC} . **a**, Average peak amplitude of I_{CRAC} ($\pm$ s.e.m., $n = 14–35$) at -80 mV induced by concentrations of InsP_3 and by the additional presence of 20 μM InsP_4 . Half-maximal effective concentration (EC_{50}) required for I_{CRAC} activation is 1.9 μM for InsP_3 and 830 nM for $\text{InsP}_3 + \text{InsP}_4$. The difference between the data sets is highly significant ($P = 0.007$, paired t -test). **b**, Average

time required for half-maximal activation of I_{CRAC} ($t_{1/2} \pm$ s.e.m., $n = 14–35$). Student's t -test evaluates the differences for data obtained at 2 μM and 3 μM InsP_3 as highly significant at $P = 0.000003$ and $P = 0.000002$, respectively. At 5 μM and 10 μM InsP_3 , the difference becomes insignificant with $P = 0.09$ and $P = 0.56$, respectively. **c**, Percentage of cells that develop I_{CRAC} .

of InsP_3 (Fig. 2b). Last, InsP_4 increases the percentage of cells that activate I_{CRAC} at low InsP_3 concentrations (Fig. 2c) so that, in the presence of InsP_4 , a dose as low as $2 \mu\text{M}$ InsP_3 can evoke I_{CRAC} in 100% of the cells tested, whereas at least $5 \mu\text{M}$ InsP_3 is needed to obtain this level of I_{CRAC} activation when only InsP_3 is used. Thus, InsP_4 is particularly effective at low InsP_3 concentrations (0.5 – $3 \mu\text{M}$), where it mediates statistically significant facilitation of I_{CRAC} activation ($P < 0.05$).

As InsP_4 can also be metabolized, albeit to a lesser extent than can InsP_3 , we investigated whether downstream metabolites of InsP_3 or InsP_4 could account for the observed stimulatory effects on I_{CRAC} . However, none of the various inositol phosphates that lack intrinsic Ca^{2+} release activity was effective in facilitating activation of I_{CRAC} at concentrations of 1 – $40 \mu\text{M}$ (data not shown). In addition, the inability of InsP_4 to induce either Ca^{2+} release or activation of I_{CRAC} (Fig. 1) eliminates the possibility that it produces inositol phosphates with release activity, such as InsP_3 or $\text{Ins}(1,4,6)\text{P}_3$ (ref. 15). Furthermore, other possible direct actions of InsP_4 do not seem to be involved in this particular aspect of InsP_4 function, as InsP_4 does not lead to activation of I_{CRAC} and seems to enhance Ca^{2+} influx only when combined with InsP_3 .

These data suggested that other InsP_3 5-phosphatase inhibitors should behave in a similar way to InsP_4 . To test this hypothesis directly, we studied three analogues that are inhibitory to InsP_3 5-phosphatase (see Fig. 3a for structures): L-*myo*-inositol 1,3,4,5-tetrakisphosphate, (L- InsP_4); the enantiomer of InsP_4 , L-threitol-1,2,4-trisphosphate (L-Thre(1,2,4) P_3); and D-2,3-bisphosphoglycerate (D-2,3-PG), which acts as a competitive inhibitor of 5-phosphatase¹⁶. We synthesized L-Thre(1,2,4) P_3 specifically to

improve the inhibitory actions of D-2,3-PG, whereas L- InsP_4 was expected to have inhibitory effects on the basis of its structural similarity to InsP_4 . We confirmed that D- InsP_4 , as well as the above compounds, inhibits 5-phosphatase by acting as a co-substrate (Fig. 3b). D- InsP_4 is the most potent 5-phosphatase inhibitor, (IC_{50} (half-maximal inhibitory concentration) $\approx 0.15 \mu\text{M}$) and is at least one order of magnitude more effective than L- InsP_4 , ($\text{IC}_{50} \approx 1.8 \mu\text{M}$). The physiological metabolite D- InsP_4 seems to be more potent than previously described synthetic 5-phosphatase inhibitors¹⁷, such as L-*chiro*- $\text{Ins}(2,3,5)\text{P}_3$ ($\text{IC}_{50} \approx 0.23 \mu\text{M}$) and L-*chiro*- $\text{Ins}(1,4,6)\text{P}_3$ ($\text{IC}_{50} \approx 0.30 \mu\text{M}$).

Figure 3e shows the average inward currents of cells perfused with the threshold concentration of $1 \mu\text{M}$ InsP_3 and each of the inhibitors. Three of the 5-phosphatase inhibitors facilitated InsP_3 -mediated activation of I_{CRAC} , with D- InsP_4 being the most effective, followed by L-Thre(1,2,4) P_3 and then D-2,3-PG (see also Table 1). Notably, L- InsP_4 was completely ineffective in facilitating I_{CRAC} , although it is a potent inhibitor of 5-phosphatase, second only to D- InsP_4 (Fig. 3b).

As D-2,3-PG interacts with InsP_3 receptors¹⁶, we looked at the potential interaction of 5-phosphatase inhibitors with InsP_3 receptors (InsP_3R). InsP_3 binding assays show that all four compounds inhibit InsP_3 -binding to InsP_3R with an order of potency similar to that of 5-phosphatase inhibition (Fig. 3c). InsP_3R binding and 5-phosphatase inhibition would have opposing effects on InsP_3 -mediated activation of I_{CRAC} , with the first contributing a positive effect and the second contributing a negative effect. The resulting net-facilitation curves (Fig. 3d) can be calculated as the product of these two factors, where the amplitude reflects efficacy and the location of the curve on the x axis reflects potency. These net-facilitation curves predict InsP_4 to be the most potent and the most effective facilitator of I_{CRAC} activation, followed by L-Thre(1,2,4) P_3 and then D-2,3-PG. L- InsP_4 is predicted to be the least effective, because its strong inhibition of InsP_3 binding essentially obliterates any facilitation through 5-phosphatase.

These predictions are confirmed qualitatively by patch-clamp experiments, which clearly show that InsP_4 provides the most effective facilitation of I_{CRAC} , whereas its enantiomer, L- InsP_4 , has no measurable effect (Fig. 3e and Table 1). D-2,3-PG and L-Thre(1,2,4) P_3 are similar both in efficacy and in delay of activation (Table 1). Although a high efficacy (y axis) of D-2,3-PG is predicted in Fig. 3d, the potency of D-2,3-PG (x axis) peaks at fairly high concentrations above $100 \mu\text{M}$; however, there is maximum enhancement of I_{CRAC} activation with 10 – $20 \mu\text{M}$ D-2,3-PG, with decreasing effects above $100 \mu\text{M}$. Similarly, the net-facilitation curve for InsP_4 predicts it to be most potent at about 100 nM to $1 \mu\text{M}$, but there is no facilitation until $5 \mu\text{M}$. Although the biochemical data are in good agreement with the rank order of efficacy

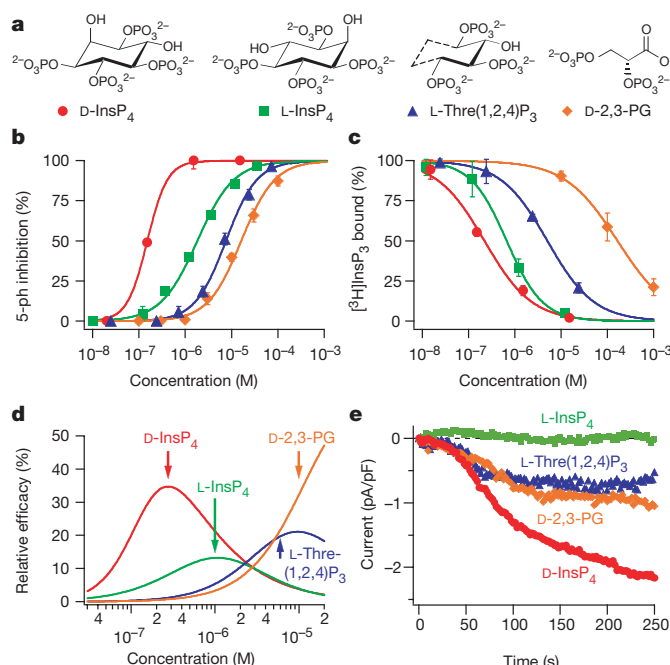


Figure 3 InsP_4 and related compounds inhibit 5-phosphatase and InsP_3 receptor binding. **a**, Chemical structures of 5-phosphatase inhibitors. L- InsP_4 is shown in a binding orientation in which it can mimic its enantiomer D- InsP_4 ; the relationship of the acyclic L-Thre(1,2,4) P_3 to InsP_3 / InsP_4 is shown by a dashed line. **b**, Inhibition profile of type I 5-phosphatase activity (see Methods). Data points are averages $\pm$ s.e.m. ($n = 3$). **c**, Inhibition profile of [^3H] InsP_3 binding to InsP_3 receptors (see Methods). Data points are averages $\pm$ s.e.m. ($n = 3$). **d**, Facilitation profiles of 5-phosphatase inhibitors. Net-facilitation curves reflect the product of inhibition profiles of 5-phosphatase activity and InsP_3 binding of InsP_3 receptors from **b** and **c**. **e**, Average inward currents carried by I_{CRAC} at -80 mV . Cells are co-perfused with various 5-phosphatase inhibitors and subthreshold levels of InsP_3 ($1 \mu\text{M}$). Traces are averages of experiments in which D- InsP_4 ($n = 14$), L- InsP_4 ($n = 15$) and L-Thre(1,2,4) P_3 ($n = 5$) were used at $20 \mu\text{M}$. D-2,3-PG ($n = 4$) was used at $10 \mu\text{M}$.

Table 1 5-Phosphatase inhibitors facilitate $\text{Ins}(1,4,5)\text{P}_3$ -mediated activation of I_{CRAC}

	D- InsP_4	L-Thre(1,2,4) P_3	D-2,3-PG	L- InsP_4
Peak current (pA/pF)*	-2.1 ± 0.4 ($n = 14$)	-1.1 ± 0.3 ($n = 5$)	-0.9 ± 0.4 ($n = 4$)	0 ($n = 15$)
Half-maximal activation (s)*	93 ± 12 ($n = 14$)	71 ± 14 ($n = 5$)	79 ± 9 ($n = 4$)	n.a.
Concentrations tested (μM)	1–20	10–100	10–1,000	1–50
Most effective concentration (μM)	5–20	100	10–20	n.a.
IC_{50} for 5-phosphatase (μM)†	0.15	7.5	16.0	1.8
IC_{50} for [^3H] $\text{Ins}(1,4,5)\text{P}_3$ binding (μM)†	0.22	5.0	165	0.66

* Values of peak currents and half-maximal activation times of I_{CRAC} represent means $\pm$ s.e.m. They are derived from time course analysis of inward currents illustrated in Fig. 3e and correspond to experiments in which inhibitors were used at $20 \mu\text{M}$, except D-2,3-PG, which was used at $10 \mu\text{M}$. † Half-maximal inhibitory concentrations (IC_{50}) for 5-phosphatase and $\text{Ins}(1,4,5)\text{P}_3$ binding are derived from dose-response fits to data sets presented in Fig. 3b and c, respectively. n.a., not applicable.

of the test compounds in facilitating I_{CRAC} , it seems that the potencies cannot be predicted accurately, simply from the combined effects of 5-phosphatase inhibition and $InsP_3$ R binding. This is not an unexpected conclusion as the predictions are based on *in vitro* biochemical analyses, which do not necessarily replicate *in vivo* conditions. In addition, the $InsP_3$ -binding data were obtained from cerebellar microsomes, which contain mainly type I $InsP_3$ receptors¹⁸, whereas RBL cells have a mixed set of $InsP_3$ receptors¹⁹, primarily comprising of type II. Possibly, interaction of $InsP_4$ and the other analogues with $InsP_3$ receptors may be subtype-specific,

just as the sensitivities of $InsP_3$ receptor subtypes to $InsP_3$ are different¹⁵.

If $InsP_4$ acts primarily by regulating $InsP_3$ metabolism, then $InsP_4$ should be largely ineffective when I_{CRAC} is activated by $Ins(2,4,5)P_3$, a non-metabolizable analogue of $InsP_3$. This is confirmed in Fig. 4a, which shows development of I_{CRAC} in cells perfused with $InsP_4$ and 5 μM or 20 μM $Ins(2,4,5)P_3$. $InsP_4$ does not facilitate $Ins(2,4,5)P_3$ -mediated activation of I_{CRAC} , but instead inhibits it; a finding that is compatible with the inhibition of Ca^{2+} influx observed in lacrimal cells²⁰ and presumably stems from the competitive antagonism of $InsP_4$ on $InsP_3$ receptors (Fig. 3c). Therefore, $InsP_4$ can compete effectively with $Ins(2,4,5)P_3$ for the $InsP_3$ receptors, but cannot influence $Ins(2,4,5)P_3$ levels as they are not regulated by 5-phosphatase.

We next tested for effects of $InsP_4$ on Ca^{2+} signals induced by receptor stimulation. Figure 4b shows carbachol-induced Ca^{2+} signals in RBL-2H3-M1 cells in the presence (5–20 μM $InsP_4$; $n = 13$) and absence of $InsP_4$. Carbachol application induces Ca^{2+} release, and the time to peak of the release transient in the presence of $InsP_4$ occurs with a statistically significant delay of 4 s (control, 4.2 ± 0.3 s; + $InsP_4$, 8.6 ± 0.9 s, $P = 0.0008$). No statistical significance was found for maximal release rates (control, 237 ± 38 nM s⁻¹; + $InsP_4$, 155 ± 28 nM s⁻¹, $P = 0.09$). This is consistent with a competitive antagonism of $InsP_4$ on $InsP_3$ receptors. After Ca^{2+} release, Ca^{2+} influx is activated and, unlike the first hyperpolarization, the second one triggers a change in $[Ca^{2+}]_i$, which, in the presence of $InsP_4$, is larger and more sustained than that of control cells. Analysis of control and $InsP_4$ -enhanced influx phases, by integrating the area under the influx transient (10 s period), reveals statistically significant enhancement of Ca^{2+} entry that leads to a twofold increase in $[Ca^{2+}]_i$ (control, 441 ± 81 nM s⁻¹; + $InsP_4$, 858 ± 109 nM s⁻¹; $P = 0.008$).

In intact cells, $InsP_4$ is produced by $InsP_3$ 3-kinases²¹ at the expense of $InsP_3$ itself. Nevertheless, $InsP_4$ could ultimately induce a net elevation of $InsP_3$ levels where metabolism through 5-phosphatase outweighs that of the 3-kinase, for example, underneath the plasma membrane where 5-phosphatase is localized²². Therefore, low levels of $InsP_3$ generated by an initial weak-agonist signal may

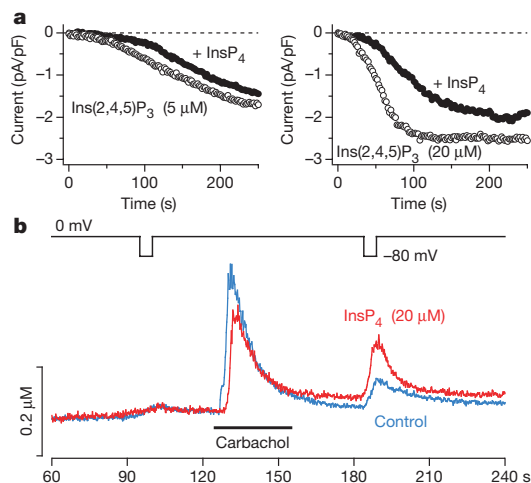


Figure 4 $InsP_4$ does not facilitate non-metabolizable $Ins(2,4,5)P_3$. **a**, Left, average inward currents carried by I_{CRAC} at -80 mV induced by 5 μM $Ins(2,4,5)P_3$ alone ($n = 8$) and by 5 μM $Ins(2,4,5)P_3$ + 20 μM $InsP_4$ ($n = 8$). Right, average inward currents induced by 20 μM $Ins(2,4,5)P_3$ ($n = 8$) and by 20 μM $Ins(2,4,5)P_3$ + 20 μM $InsP_4$ ($n = 7$). **b**, Average $[Ca^{2+}]_i$ signals in patch-clamped RBL-2H3-M1 cells. Blue control trace shows cells perfused with standard internal solution ($n = 10$); red trace shows responses of cells perfused with 20 μM $InsP_4$ ($n = 8$). Identical hyperpolarizations to -80 mV were induced before and after carbachol (100 μM) challenge to increase the driving force for Ca^{2+} and probe the magnitude of Ca^{2+} entry.

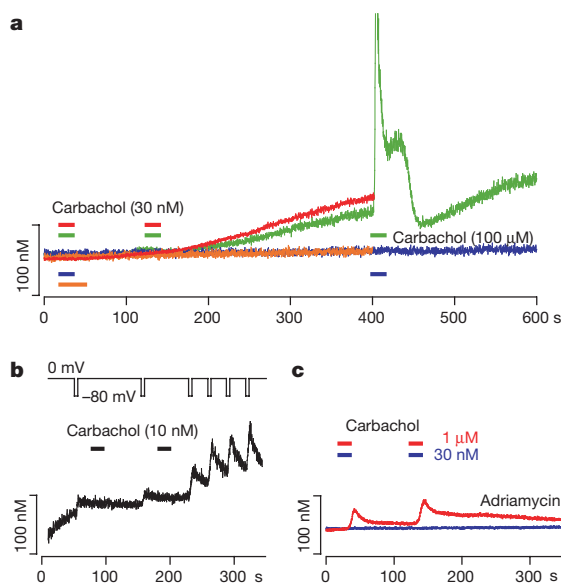


Figure 5 $InsP_4$ selectively facilitates Ca^{2+} influx. **a**, Average intracellular Ca^{2+} signals in intact RBL-2H3-M1 cells stimulated by low levels of carbachol. Two identical carbachol stimuli (30 nM) of 15-s duration were delivered at intervals of 0 s (orange trace, $n = 7$), 90 s (red trace, $n = 20$) and 365 s (blue trace, $n = 6$). The green trace (shifted down by 5 nM for illustration purposes) represents a subset of 6 cells of the 20 red trace cells that were stimulated a third time by 100 μM carbachol in Ca^{2+} -free solution. **b**, Average

intracellular Ca^{2+} signals ($n = 3$) of cells perfused with standard internal solution and stimulated as in **a**. **c**, Average intracellular Ca^{2+} signals in intact RBL-2H3-M1 cells stimulated by subthreshold (30 nM, $n = 14$) and threshold (1 μM , $n = 6$) levels of carbachol. Stimulation of cells was performed as in **a** after pre-incubation with adriamycin (10 μM) for 2 h. The red trace was shifted up by 10 nM for display purposes.

induce spatially confined InsP_3 production that, owing to effective metabolism, is not sufficient to cause Ca^{2+} release or to activate I_{CRAC} . However, the concomitant production of InsP_4 might prime and sensitize cells to respond more effectively to a second stimulation by the same, or different agonist. We challenged receptor-mediated Ca^{2+} signals in intact RBL-2H3-M1 cells with two identical stimuli of 30 nM carbachol, which were delivered for 15 s at different time intervals (Fig. 5a). This agonist concentration is well below the threshold for measurable Ca^{2+} release, and we observed no visible Ca^{2+} signal after the first stimulus. Continued stimulation for another 15 s also failed to induce a Ca^{2+} signal (Fig. 5a; orange trace). However, when the second identical stimulus was given after 90 s, 20 out of 24 cells generated a long-lasting plateau of elevated $[\text{Ca}^{2+}]_i$ (Fig. 5a, red trace). Such facilitation was not observed when the second carbachol application was delivered with a delay of 365 s (Fig. 5a, blue trace).

These results suggest that the first carbachol stimulation generates a priming factor that facilitates subsequent Ca^{2+} signalling. In RBL-2H3-M1 cells, carbachol only produces diacylglycerol and InsP_3 . As diacylglycerol activates protein kinase C, which is known to cause inhibition of Ca^{2+} influx in RBL cells²³, we must conclude that the facilitation is due to an inositol phosphate, specifically InsP_4 , as its lifetime is consistent with the time window of facilitation²⁴ and all other inositol phosphates failed to facilitate I_{CRAC} (data not shown). The results also suggest that a local increase in InsP_3 levels can selectively release Ca^{2+} from stores that activate I_{CRAC} ('CRAC stores'). As Ca^{2+} release from these 'CRAC stores' does not contribute significantly to global cytosolic Ca^{2+} signals¹⁰, there is only a Ca^{2+} influx signal and no visible Ca^{2+} transient. The experiments shown by the green trace in Fig. 5a support this idea. There the facilitation of Ca^{2+} signalling proceeds as in the red trace; a subsequent challenge with a large dose of carbachol (100 μM) releases all Ca^{2+} from InsP_3 -sensitive stores, which results in a large Ca^{2+} release transient. Thus, it seems that 'CRAC stores' can be emptied by localized InsP_3 signalling, whereas the bulk of InsP_3 -sensitive stores remain full.

The experiments in intact cells were done in Cs^+ -containing external solution, which minimizes agonist-mediated changes in membrane potential that could indirectly enhance Ca^{2+} influx. To further exclude any such effect and to ascertain that the Ca^{2+} signal facilitation was due to Ca^{2+} influx, we confirmed that facilitation of a second subthreshold carbachol stimulus also occurs when stimulating cells in the whole-cell configuration (Fig. 5b). As in the intact cells, carbachol did not induce Ca^{2+} release, but selectively facilitated Ca^{2+} influx, as changes in $[\text{Ca}^{2+}]_i$ during membrane hyperpolarizations were significantly enhanced after the second, but not the first, application of carbachol.

We also tested adriamycin (doxorubicin), which is known to inhibit InsP_3 3-kinase²⁵ and suppress InsP_4 production. Cells treated with 10 μM adriamycin did not have a facilitated Ca^{2+} influx component after a second subthreshold carbachol stimulation (Fig. 5c, blue trace). Adriamycin also suppressed facilitation in patch-clamped cells and did not inhibit I_{CRAC} directly (data not shown), thereby eliminating nonspecific drug effects on ion channels. We can also eliminate reduced InsP_3 production as a possible side-effect, as the threshold carbachol concentration of 1 μM in adriamycin-treated cells was similar to that of control RBL cells and produced enough InsP_3 to consistently reach the threshold for visible Ca^{2+} release (Fig. 5c, red trace). Even under these conditions, adriamycin largely suppressed the facilitatory effects on Ca^{2+} signalling. Therefore, the adriamycin experiments are consistent with the hypothesis that InsP_4 production mediated by InsP_3 3-kinase is primarily responsible for the observed facilitation of repetitive subthreshold stimuli.

We have presented evidence that shows InsP_4 to be a potent, naturally occurring 5-phosphatase inhibitor that facilitates Ca^{2+} influx by sensitizing InsP_3 -mediated activation of I_{CRAC} . As InsP_4

production occurs at the expense of InsP_3 , and high levels of InsP_4 can inhibit InsP_3 receptors, it is possible that in some cell systems, the overall effect of InsP_4 on Ca^{2+} signalling may be inhibitory. This would apply particularly to cells with high levels of endogenous or experimentally over-expressed InsP_3 3-kinase^{22,26}. The impact of InsP_4 on Ca^{2+} signalling may be multifaceted and cell-type specific, varying with strength of receptor stimulation, subcellular localization of InsP_3 -metabolizing enzymes, heterogeneity of intracellular Ca^{2+} stores, InsP_3 receptor subtype composition and additional regulatory mechanisms that affect the above. □

Methods

Synthetic inositol phosphates and 5-phosphatase inhibitors

We synthesized D- InsP_4 and L- InsP_4 as described²⁷. L-Thre(1,2,4)P₃ was prepared from (2S,3S)-(+)-2-benzyloxybutane-1,3,4-triol (Fluka). Briefly, phosphorylation of the triol with N,N-diethyl-1,5-dihydro-2,4,3-benzodioxaphosphin-3-amine in the presence of 1H-tetrazole, followed by oxidation *in situ* with 3-chloroperoxybenzoic acid, gave the protected trisphosphate, which was purified by flash chromatography on silica gel. Deprotection by hydrogenolysis over palladium on carbon gave L-Thre(1,2,4)P₃, which was purified by ion-exchange chromatography on Q-Sepharose fast flow resin before use. All synthetic compounds showed analytical and spectroscopic data in full accordance with structure. We obtained other inositol phosphates and D-2,3-PG from Sigma.

5-Phosphatase inhibition and InsP_3 receptor-binding assays

The assays of InsP_3 5-phosphatase activity were performed at pH 7.2 using purified recombinant enzyme expressed in *Escherichia coli* as described²⁸. We performed the InsP_3 receptor-binding assays by using microsomal fractions of rat cerebellum prepared as described²⁹. The assay mixture (0.5 ml) contained 50 mM HEPES/NaOH buffer pH 7.2, 1 mM EDTA, 6 mM [3H] InsP_3 and ~30 μg of microsomal fraction, in the presence of various concentrations of compounds of interest. Incubation was performed on ice for 10 min, followed by the separation of bound radioactivity from free form by centrifugation (15,000 r.p.m. for 10 min). Nonspecific binding (150–200 d.p.m.) was determined in the presence of 10 μM InsP_3 and was subtracted from that in its absence to determine the specific binding (4,000–5,000 d.p.m.).

Electrophysiology

For patch-clamp experiments, RBL-2H3-M1 cells grown on glass coverslips were transferred to the recording chamber and kept in a standard modified Ringer's solution containing (in mM): 145 NaCl, 2.8 KCl, 10 CsCl, 10 CaCl_2 , 2 MgCl_2 , 10 glucose, 10 HEPES NaOH, pH 7.2. We used CsCl to inhibit inward rectifier potassium currents. For Ca^{2+} measurements, the external Ca^{2+} concentration was adjusted to 2 mM. We used carbachol at the indicated concentrations. The standard intracellular pipette-filling solution contained (in mM): 145 caesium-glutamate, 8 NaCl, 1 MgCl_2 , 0.5 Mg-ATP , 0.3 GTP, pH 7.2, adjusted with CsOH. Except for the fura-2 experiments, the internal solution was supplemented with a mixture of 10 mM caesium-BAPTA and 4.3–5.3 mM CaCl_2 to buffer $[\text{Ca}^{2+}]_i$ to resting levels of 100–150 nM and to avoid spontaneous activation of I_{CRAC} . We carried out patch-clamp experiments in the tight-seal whole-cell configuration at 21–25 °C. We acquired high-resolution current recordings by a computer-based patch-clamp amplifier system (EPC-9, HEKA, Lambrecht, Germany). Patch pipettes had resistances between 2 and 4 M Ω after filling with the standard intracellular solution. Immediately after establishment of the whole-cell configuration, voltage ramps of 50-ms duration spanning the voltage range of –100 to +100 mV were delivered from a holding potential of 0 mV at a rate of 0.5 Hz over a period of 300–400 s. All voltages were corrected for a liquid junction potential of 10 mV between external and internal solutions. We filtered currents at 2.3 kHz and digitized them at 100- μs intervals. Capacitive currents and series resistance were determined and corrected before each voltage ramp using the automatic capacitance compensation of the EPC-9. For analysis, the very first ramps before activation of I_{CRAC} (usually 1–3) were digitally filtered at 2 kHz, pooled and used for leak subtraction of all subsequent current records. The low-resolution temporal development of inward currents was extracted from the leak-corrected individual ramp current records by measuring the current amplitude at –80 mV.

Calcium measurements

The cytosolic calcium concentration of individual patch-clamped or intact cells was monitored at a rate of 5 Hz with a photomultiplier-based system using a monochromatic light source tuned to excite fura-2 fluorescence at 360 and at 390 nm for 20 ms each. Emission was detected at 450–550 nm with a photomultiplier whose analogue signals were sampled and processed by the X-Chart software package (HEKA, Lambrecht, Germany). Fluorescence ratios were translated into free intracellular calcium concentration on the basis of calibration parameters derived from patch-clamp experiments with calibrated calcium concentrations. In patch-clamp experiments, we added fura-2 to the standard intracellular solution at 100 μM . Ester loading of intact cells was performed by incubating cells for 45–60 min in standard solution (2 mM extracellular calcium) supplemented with 5 μM fura-2-AM. In all experiments, where intracellular Ca^{2+} was monitored (patch-clamp or intact cells), the standard external solution contained 2 mM Ca^{2+} . Local perfusion of individual cells with carbachol was achieved through a wide-tipped, pressure-controlled application pipette (3 μm diameter) placed approximately 30 μm from the cell under investigation.

Received 19 July; accepted 20 September 2000.

- Parekh, A. B. & Penner, R. Store depletion and calcium influx. *Physiol. Rev.* **77**, 901–930 (1997).
- Connolly, T. M., Bansal, V. S., Bross, T. E., Irvine, R. F. & Majerus, P. W. The metabolism of tris- and tetraphosphates of inositol by 5-phosphomonoesterase and 3-kinase enzymes. *J. Biol. Chem.* **262**, 2146–2149 (1987).
- Irvine, R. F. & Moor, R. M. Micro-injection of inositol 1,3,4,5-tetrakisphosphate activates sea urchin eggs by a mechanism dependent on external Ca^{2+} . *Biochem. J.* **240**, 917–920 (1986).
- Morris, A. P., Gallacher, D. V., Irvine, R. F. & Petersen, O. H. Synergism of inositol trisphosphate and tetrakisphosphate in activating Ca^{2+} -dependent K^{+} channels. *Nature* **330**, 653–655 (1987).
- Bird, G. S. *et al.* Activation of Ca^{2+} entry into acinar cells by a non-phosphorylatable inositol trisphosphate. *Nature* **352**, 162–165 (1991).
- Smith, P. M., Harmer, A. R., Letcher, A. J. & Irvine, R. F. The effect of inositol 1,3,4,5-tetrakisphosphate on inositol trisphosphate-induced Ca^{2+} mobilization in freshly isolated and cultured mouse lacrimal acinar cells. *Biochem. J.* **347**, 77–82 (2000).
- Fukuda, M. & Mikoshiba, K. The function of inositol high polyphosphate binding proteins. *BioEssays* **19**, 593–603 (1997).
- Cullen, P. J. Bridging the GAP in inositol 1,3,4,5-tetrakisphosphate signalling. *Biochim. Biophys. Acta* **1436**, 35–47 (1998).
- Hoth, M. & Penner, R. Depletion of intracellular calcium stores activates a calcium current in mast cells. *Nature* **355**, 353–356 (1992).
- Parekh, A. B., Fleig, A. & Penner, R. The store-operated calcium current I_{CRAC} : nonlinear activation by InsP_3 and dissociation from calcium release. *Cell* **89**, 973–980 (1997).
- Broad, L. M., Armstrong, D. L. & Putney, J. W. Role of the inositol 1,4,5-trisphosphate receptor in Ca^{2+} feedback inhibition of calcium release-activated calcium current I_{CRAC} . *J. Biol. Chem.* **274**, 32881–32888 (1999).
- Glitsch, M. D. & Parekh, A. B. Ca^{2+} store dynamics determines the pattern of activation of the store-operated Ca^{2+} current I_{CRAC} in response to InsP_3 in rat basophilic leukaemia cells. *J. Physiol. (Lond.)* **523**, 283–290 (2000).
- Jones, S. V., Choi, O. H. & Beaven, M. A. Carbachol induces secretion in a mast cell line (RBL-2H3) transfected with the M1 muscarinic receptor gene. *FEBS Lett.* **289**, 47–50 (1991).
- Lewis, R. S. & Cahalan, M. D. Potassium and calcium channels in lymphocytes. *Annu. Rev. Immunol.* **13**, 623–653 (1995).
- Hirata, M. *et al.* Inositol 1,4,5-trisphosphate receptor subtypes differentially recognize regioisomers of *D*-myo-inositol 1,4,5-trisphosphate. *Biochem. J.* **328**, 93–98 (1997).
- Guillemette, G., Favreau, L., Lamontagne, S. & Boulay, G. 2,3-Diphosphoglycerate is a nonselective inhibitor of inositol 1,4,5-trisphosphate action and metabolism. *Eur. J. Pharmacol.* **188**, 251–260 (1990).
- Safrany, S. T. *et al.* Design of potent and selective inhibitors of myo-inositol 1,4,5-trisphosphate 5-phosphatase. *Biochemistry* **33**, 10763–10769 (1994).
- Wojcikiewicz, R. J. Type I, II, and III inositol 1,4,5-trisphosphate receptors are unequally susceptible to down-regulation and are expressed in markedly different proportions in different cell types. *J. Biol. Chem.* **270**, 11678–11683 (1995).
- De Smedt, H. *et al.* Determination of relative amounts of inositol trisphosphate receptor mRNA isoforms by ratio polymerase chain reaction. *J. Biol. Chem.* **269**, 21691–21698 (1994).
- Bird, G. S. & Putney, J. W. Jr. Effect of inositol 1,3,4,5-tetrakisphosphate on inositol trisphosphate-activated Ca^{2+} signaling in mouse lacrimal acinar cells. *J. Biol. Chem.* **271**, 6766–6770 (1996).
- Communi, D., Dewaste, V. & Erneux, C. Calcium-calmodulin-dependent protein kinase II and protein kinase C-mediated phosphorylation and activation of *D*-myo-inositol 1,4,5-trisphosphate 3-kinase B in astrocytes. *J. Biol. Chem.* **274**, 14734–14742 (1999).
- De Smedt, F. *et al.* Isoprenylated human brain type I inositol 1,4,5-trisphosphate 5-phosphatase controls Ca^{2+} oscillations induced by ATP in Chinese hamster ovary cells. *J. Biol. Chem.* **272**, 17367–17375 (1997).
- Parekh, A. B. & Penner, R. Depletion-activated calcium current is inhibited by protein kinase in RBL-2H3 cells. *Proc. Natl Acad. Sci. USA* **92**, 7907–7911 (1995).
- Hughes, A. R., Takemura, H. & Putney, J. W. Jr. Kinetics of inositol 1,4,5-trisphosphate and inositol cyclic 1:2,4,5-trisphosphate metabolism in intact rat parotid acinar cells. Relationship to calcium signalling. *J. Biol. Chem.* **263**, 10314–10319 (1988).
- da Silva, C. P., Emmrich, F. & Guse, A. H. Adriamycin inhibits inositol 1,4,5-trisphosphate 3-kinase activity *in vitro* and blocks formation of inositol 1,3,4,5-tetrakisphosphate in stimulated Jurkat T-lymphocytes. Does inositol 1,3,4,5-tetrakisphosphate play a role in Ca^{2+} -entry? *J. Biol. Chem.* **269**, 12521–12526 (1994).
- Balla, T. *et al.* Agonist-induced calcium signaling is impaired in fibroblasts overproducing inositol 1,3,4,5-tetrakisphosphate. *J. Biol. Chem.* **266**, 24719–24726 (1991).
- Riley, A. M., Mahon, M. F. & Potter, B. V. L. Rapid synthesis of the enantiomers of myo-inositol 1,3,4,5-tetrakisphosphate by direct chiral desymmetrization of myo-inositol orthoformate. *Angew. Chem. Int. Edn. Eng.* **36**, 1472–1474 (1997).
- Yoshimura, K., Watanabe, Y., Erneux, C. & Hirata, M. Use of phosphorofluoridate analogues of *D*-myo-inositol 1,4,5-trisphosphate to assess the involvement of ionic interactions in its recognition by the receptor and metabolising enzymes. *Cell. Signal.* **11**, 117–125 (1999).
- Worley, P. F., Baraban, J. M., Supattapone, S., Wilson, V. S. & Snyder, S. H. Characterization of inositol trisphosphate receptor binding in brain. Regulation by pH and calcium. *J. Biol. Chem.* **262**, 12132–12136 (1987).

Acknowledgements

We would like to thank D. Tani and M. Monteilh-Zoller for technical assistance; C. Erneux for the InsP_3 5-phosphatase plasmid ECH10; the Wellcome Trust for Programme Grant Support (to B.V.L.P.). We acknowledge grant support by the Ministry of Education, Science, Sports and Culture of Japan (to H.T. and M.H.), Kyushu University Interdisciplinary Programs in Education and Projects in Research Development (to M.H.) and The Naito Foundation (to M.H.).

Correspondence should be addressed to R.P. (e-mail: rpenner@hawaii.edu) and requests for materials should be addressed to B.V.L.P. (e-mail: prsbvlp@bath.ac.uk).

A Toll-like receptor recognizes bacterial DNA

Hiroaki Hemmi^{*†}, Osamu Takeuchi^{*†}, Taro Kawai^{*†}, Tsuneyasu Kaisho^{*†}, Shintaro Sato^{*†}, Hideki Sanjo^{*†}, Makoto Matsumoto^{*†}, Katsuaki Hoshino^{*†}, Hermann Wagner[‡], Kiyoshi Takeda^{*†} & Shizuo Akira^{*†}

^{*} Department of Host Defense, Research Institute for Microbial Diseases, Osaka University and [†] Core Research for Evolutional Science and Technology, Japan Science and Technology Corporation, 3-1 Yamada-oka, Suita, Osaka 565-0871, Japan

[‡] Institute of Medical Microbiology, Immunology and Hygiene, Technical University of Munich, Trogerstr. 9, D-81675 Munich, Germany

DNA from bacteria has stimulatory effects on mammalian immune cells^{1–3}, which depend on the presence of unmethylated CpG dinucleotides in the bacterial DNA. In contrast, mammalian DNA has a low frequency of CpG dinucleotides, and these are mostly methylated; therefore, mammalian DNA does not have immuno-stimulatory activity. CpG DNA induces a strong T-helper-1-like inflammatory response^{4–7}. Accumulating evidence has revealed the therapeutic potential of CpG DNA as adjuvants for vaccination strategies for cancer, allergy and infectious diseases^{8–10}. Despite its promising clinical use, the molecular mechanism by which CpG DNA activates immune cells remains unclear. Here we show that cellular response to CpG DNA is mediated by a Toll-like receptor, TLR9. TLR9-deficient (TLR9^{−/−}) mice did not show any response to CpG DNA, including proliferation of splenocytes, inflammatory cytokine production from macrophages and maturation of dendritic cells. TLR9^{−/−} mice showed resistance to the lethal effect of CpG DNA without any elevation of serum pro-inflammatory cytokine levels. The *in vivo* CpG-DNA-mediated T-helper type-1 response was also abolished in TLR9^{−/−} mice. Thus, vertebrate immune systems appear to have evolved a specific Toll-like receptor that distinguishes bacterial DNA from self-DNA.

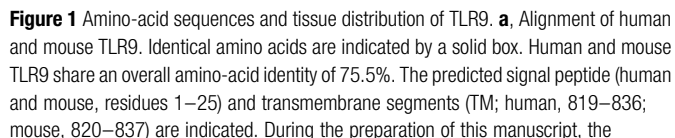
The Toll-like receptor (TLR) family is a phylogenetically conserved mediator of innate immunity that is essential for microbial recognition¹¹. Mammalian TLRs comprise a large family with extracellular leucine-rich repeats (LRRs) and a cytoplasmic Toll/interleukin (IL)-1R (TIR) homology domain. So far, six members (TLR1–6) have been reported^{12–14}, and two additional members have been deposited in GenBank as TLR7 and TLR8 (accession numbers AF240467 and AF246971, respectively). TLR2 and TLR4 are responsible for immune responses to peptidoglycan (PGN) and lipopolysaccharide (LPS), respectively^{15–22}.

By using a BLAST search, we identified an expressed sequence tag (EST) clone (AA273731; mouse) that showed high similarity with the previously identified TLRs. Using this fragment as a probe, we isolated a full-length complementary DNA from the mouse macrophage cDNA library. We also isolated the human counterpart. Sequence analysis revealed the presence of regions conserved in the TLR family, such as LRR and TIR domain (Fig. 1a, b). Therefore, we designated this gene TLR9. Northern blot analysis of various tissues indicated that mouse TLR9 transcripts were most abundantly expressed in the spleen (Fig. 1c).

To assess the biological function of TLR9, we generated TLR9^{−/−} mice by homologous recombination in embryonic stem (ES) cells. The targeting vector was constructed to replace a 1.0-kb fragment of the mouse *Tlr9* gene encoding a part of LRR with a neomycin resistance cassette (*neo*) (Fig. 2a). Correctly targeted ES cell clones were micro-injected into C57BL/6 blastocysts, which contributed to transmission of the mutated allele through the germ line. We intercrossed heterozygotes to produce offspring that were homozygous for the disrupted *Tlr9* allele (Fig. 2b). The mutant mice were

MyD88 is an adaptor molecule involved in the signalling through the IL-1R and TLR families. We previously showed that MyD88 is essential for the response to IL-1, IL-18, LPS and many other

bacterial cell-wall components²³. We have also shown that the responses to CpG DNA are dependent on MyD88 and TRAF6 (ref. 24). MyD88^{-/-} mice did not respond to CpG DNA, whereas both TLR2^{-/-} and TLR4^{-/-} mice responded normally to CpG DNA. These data indicate that CpG DNA is recognized by TLRs other than TLR2 and TLR4. Therefore, we analysed responses of TLR9^{-/-} mice to CpG DNA. We first investigated the proliferation of splenocytes in response to CpG DNA (Fig. 3a). CpG DNA, but not non-CpG DNA induced proliferation of wild-type splenocytes in a dose-dependent manner. In contrast, TLR9^{-/-} splenocytes did not proliferate in response to either CpG DNA or non-CpG DNA, although they showed a similar proliferative response to LPS as the wild-type cells. Wild-type B cells showed enhanced surface expression of major histocompatibility complex (MHC) class II in response to CpG DNA; however, a CpG-DNA-induced increase in MHC class II expression was not observed in TLR9^{-/-} B cells (data not shown). These data indicate that TLR9^{-/-} B cells are defective in



sequences for human TLR9 were deposited in GenBank by two other groups (accession numbers NM017442 and AF245704). **b**, Alignment of the cytoplasmic domains of human TLR family members. Amino-acid residues conserved in at least four molecules are highlighted by solid boxes. **c**, A mouse multiple-tissue northern blot (Clontech) containing 2 μ g of poly(A)⁺ RNA was probed with a mouse TLR9 cDNA fragment.

their response to CpG DNA.

Next, we measured production of inflammatory cytokines from peritoneal macrophages by enzyme-linked immunoabsorbent assay (ELISA; Fig. 3b). Macrophages from wild-type mice produced tumour necrosis factor (TNF)- α , IL-6 and IL-12 in response to CpG DNA. The production was further enhanced when stimulated with a combination of interferon (IFN)- γ and CpG DNA. However, macrophages from TLR9^{-/-} mice did not produce any detectable levels of inflammatory cytokines in response to CpG DNA even in the presence of IFN- γ . Macrophages from wild-type and TLR9^{-/-} mice produced similar amounts of TNF- α , IL-6 and IL-12 in response to LPS, PGN, lipoprotein from *Escherichia coli*, Zymosan and whole heat-killed *Staphylococcus aureus* (Fig. 3b; and data not shown), indicating that TLR9^{-/-} macrophages are specifically defective in their response to CpG DNA.

CpG-containing bacterial DNA is a potent stimulant for dendritic cells (DCs) to support T-helper type-1 (Th1) cell development⁴⁻⁷. Therefore, we examined CpG-DNA-induced cytokine production and upregulation of surface molecules in bone-marrow-derived murine DCs. Wild-type DCs produced IL-12 in response to CpG DNA; however, TLR9^{-/-} DCs did not produce any detectable levels of IL-12 (Fig. 3c). Wild-type DCs showed enhanced surface expression of CD40, CD80, CD86 and MHC class II when stimulated with CpG DNA. TLR9^{-/-} DCs did not show any enhanced expression of the surface molecules in response to CpG-DNA (Fig. 3d). Both wild-type and TLR9^{-/-} DCs exhibited similar responses to LPS. Together, these findings indicate that TLR9 is essential for cellular responses to CpG DNA.

Signalling through TLRs occurs through the sequential

recruitment of the adaptor molecule MyD88 and the serine/threonine kinase IRAK, and subsequently activates mitogen-activated protein (MAP) kinases and the nuclear factor NF- κ B²³. We next analysed activation of the intracellular signalling cascade in response to CpG DNA. In wild-type macrophages, stimulation with CpG DNA increased the DNA-binding activity of NF- κ B, as determined by electrophoretic mobility shift assay (EMSA; Fig. 3e); however, NF- κ B activity was not increased in response to CpG DNA in TLR9^{-/-} macrophages. LPS stimulation of TLR9^{-/-} macrophages led to activation of NF- κ B to the same extent as that of wild-type cells, indicating that CpG-DNA-induced activation of NF- κ B was impaired in TLR9^{-/-} macrophages. *In vitro* kinase assay showed that CpG DNA activated c-Jun N-terminal kinase (JNK) and IRAK in wild-type macrophages. Activation of both kinases was completely abolished in TLR9^{-/-} macrophages (Fig. 3f, g). Thus, CpG-DNA-mediated signal transduction is dependent on TLR9.

Finally, we addressed the *in vivo* response to CpG DNA in TLR9^{-/-} mice. CpG DNA can induce lethal shock in D-galactosamine (D-GalN)-sensitized mice²⁵. Wild-type mice died within 12 h after D-GalN plus CpG DNA administration with marked elevation of serum concentrations of TNF- α , IL-6 and IL-12 (Fig. 4a, b). In contrast, all TLR9^{-/-} mice survived without any increase in serum concentration of these inflammatory cytokines. Thus, TLR9^{-/-} mice were highly resistant to CpG-DNA-induced shock syndrome. *In vivo* administration of CpG DNA has also been shown to induce a Th1-biased response²⁶. CpG DNA and ovalbumin (OVA) were injected into the footpads, and lymph node cells were isolated at the 7-day time point and stimulated with OVA. The popliteal lymph node of

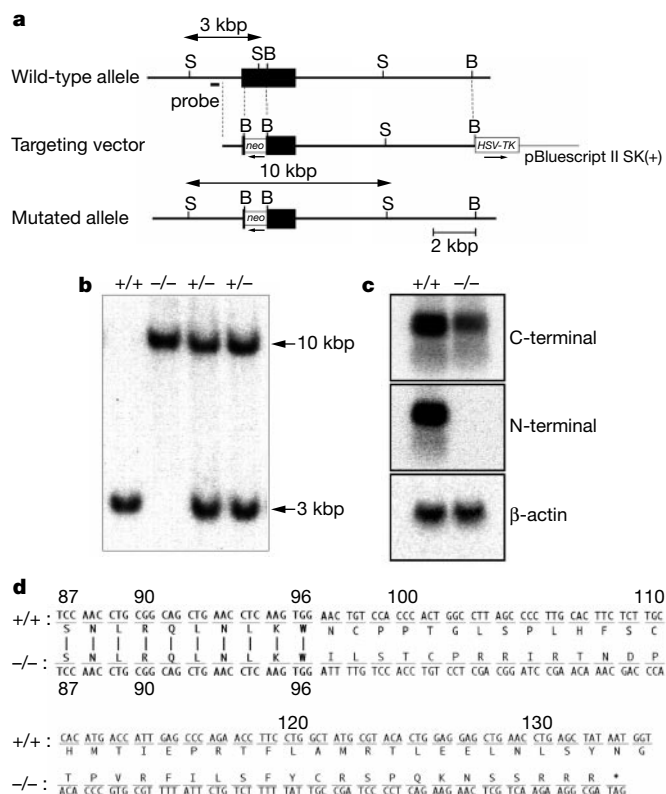


Figure 2 Targeted disruption of the mouse *Tlr9* gene. **a**, Maps of the TLR9 genome, the targeting vector and the predicted disrupted gene. Filled boxes denote the coding exon. Restriction enzymes: B, *Bam*HI; S, *Sca*I. **b**, Southern blot analysis of offspring from the heterozygote intercrosses. Genomic DNA was extracted from mouse tails, digested with *Sca*I, electrophoresed and hybridized with the radiolabelled probe indicated in **a**. Southern blotting gave a single 3.0-kb band for wild-type (+/+), a 10-kb band for homozygous

(-/-) and both bands for heterozygous mice (+/-). **c**, Northern blot analysis of splenocytes. Total RNA (10 μ g) extracted from splenocytes was electrophoresed, transferred to a nylon membrane, and hybridized using the TLR9 C-terminal or N-terminal fragment as a probe. The same membrane was rehybridized with a β -actin probe. **d**, Comparison of predicted amino-acid sequences between wild-type (+/+) and TLR9^{-/-} (-/-) cDNA. The numbers indicate predicted amino-acid position of wild-type TLR9.

CpG DNA-treated wild-type mice showed an increase in size compared with that of PBS treated mice, whereas TLR9^{-/-} mice did not show any lymphadenopathy (data not shown). Lymph node cells from CpG-DNA-treated wild-type mice produced IFN- γ in response to OVA (Fig. 4c). In contrast, production of IFN- γ from TLR9^{-/-} lymph node cells was not observed. Thus, CpG-DNA-induced Th1-like response was not observed in TLR9^{-/-} mice.

The nature and localization of the CpG receptor are controversial. There is evidence that CpG DNA binds to cell-surface receptors that subsequently transduce stimulatory signals, because Sepharose beads coated with active CpG DNA stimulate B cells as free CpG DNA²⁷. In contrast, other reports show that internalization of the DNA is required for activity²⁸. Inhibitors of endosomal maturation such as bafilomycin A or chloroquine abolish CpG-mediated cell activation, indicating that cellular uptake by nonspecific endocytosis and subsequent endosomal maturation precede cell activation^{29,30}. Acidification of endosomal CpG DNA is coupled to the rapid generation of intracellular reactive oxygen species, followed by NF- κ B activation³¹. Thus, in the latter case, it has been

proposed that CpG DNA works through binding to an intracellular receptor. The presence of a transmembrane segment in the TLR9 gene strongly suggests that TLR9 is inserted into the membrane, but not present in the cytoplasm. Although the localization of TLR9 awaits assessment by immunostaining, confocal data show that tagged MyD88 colocalizes with tagged CpG DNA in endosomal structures, but not at the cell membrane (H.W., unpublished data). In contrast, LPS colocalizes with MyD88 at the cell membrane. This suggests that signalling is triggered by LPS at the cell membrane, whereas CpG DNA initiates signalling after translocation to endosomes. This assumption may well correlate with the finding that CpG-DNA-induced IRAK activation is delayed as compared with that stimulated with LPS (Fig. 3g). The identification of CpG DNA signalling receptor will pave the way to understanding the mechanism by which CpG DNA is recognized as well as by which the recognition of CpG DNA is translated into a strong Th1 response. Furthermore, TLR9^{-/-} mice will provide a useful model for clarifying to what extent recognition of CpG DNA contributes to host immune responses against bacterial infections. □

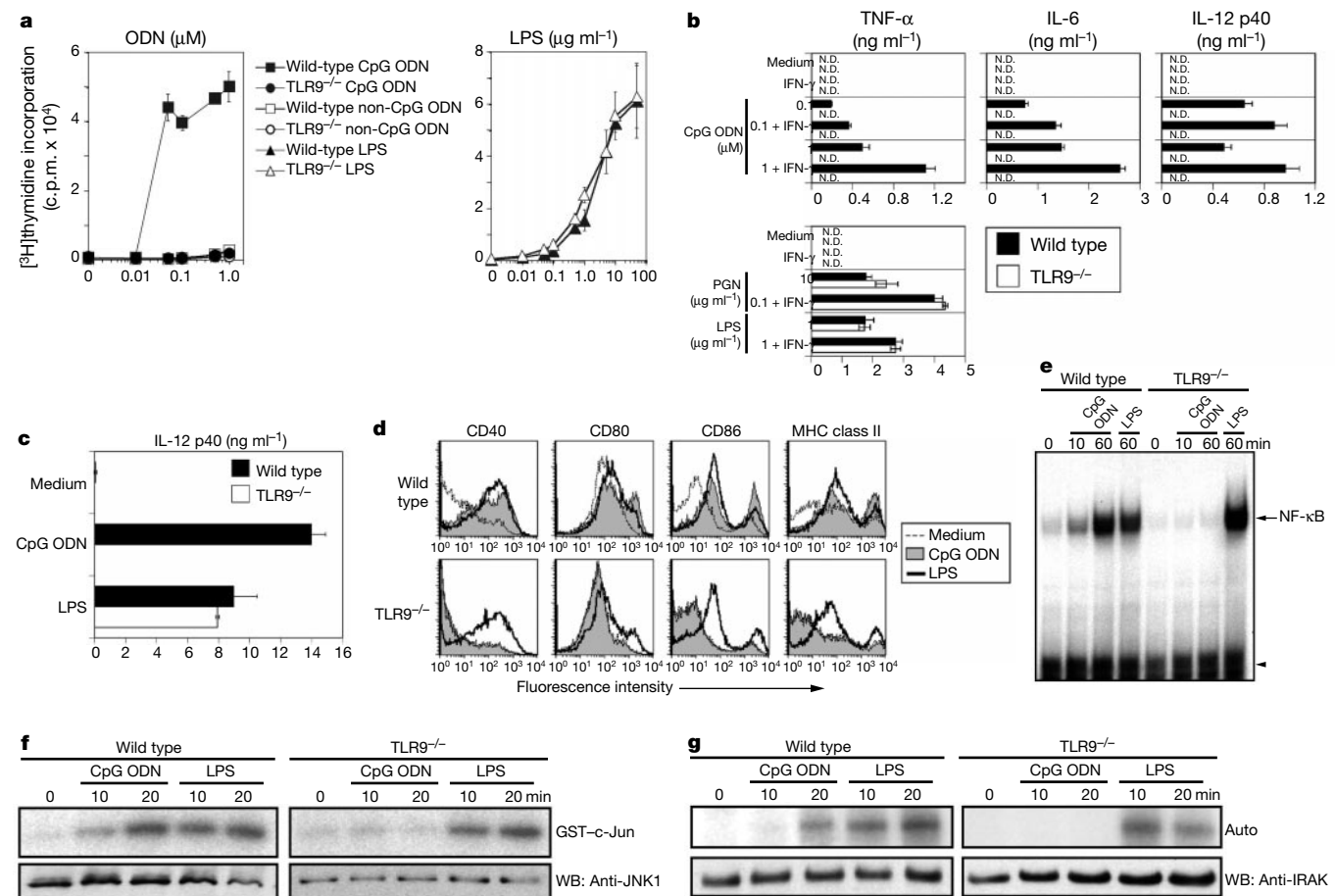


Figure 3 Impaired responses to CpG ODN in TLR9^{-/-} cells. **a**, Splenocytes from wild-type or TLR9^{-/-} mice were cultured with the indicated concentrations of CpG ODN, non-CpG ODN or LPS for 48 h plus pulsed [³H]thymidine for the last 8 h. [³H]thymidine incorporation was measured by β -scintillation counting. Data indicate mean $\pm$ s.d. of triplicate samples of one representative experiment. **b**, Peritoneal macrophages from wild-type or TLR9^{-/-} mice were stimulated with CpG ODN (0.1 or 1.0 μ M), PGN (10 μ g ml⁻¹) or LPS (1.0 μ g ml⁻¹) in the presence or absence of 30 U ml⁻¹ IFN- γ for 24 h, and concentrations of TNF- α , IL-6 and IL-12 p40 in the culture supernatants were measured by ELISA. Similar results were obtained from three independent experiments. Data indicate mean $\pm$ s.d. N.D., not detected. **c**, Wild-type or TLR9^{-/-} dendritic cells (DCs) derived from bone marrow were cultured with CpG ODN or LPS. Concentrations of IL-12 p40 in culture supernatants were measured by ELISA. Data indicate mean $\pm$ s.d. **d**, DCs were stimulated with CpG ODN or

LPS for 48 h and analysed for cell-surface expression of the indicated molecules by flow cytometry. **e**, Peritoneal macrophages from wild-type and TLR9^{-/-} mice were stimulated with 1.0 μ M CpG ODN or 1.0 μ g ml⁻¹ LPS for the indicated durations. NF- κ B activity was determined by EMSA. Arrow indicates the inducible NF- κ B complex; arrowhead indicates free probe. **f**, Peritoneal macrophages from wild-type and TLR9^{-/-} mice were stimulated with 1.0 μ M CpG ODN or 1.0 μ g ml⁻¹ LPS for the indicated durations. Cell lysates were prepared and immunoprecipitated with anti-JNK antibody. JNK activity was measured by *in vitro* kinase assay using a GST-c-Jun fusion as a substrate (top). The same lysates were blotted with anti-JNK (bottom). **g**, The same cell lysates in **f** were immunoprecipitated with anti-IRAK antibody. The kinase activity of IRAK was measured by *in vitro* kinase assay (top). The same lysates were blotted with anti-IRAK antibody (bottom). Auto, autophosphorylation.

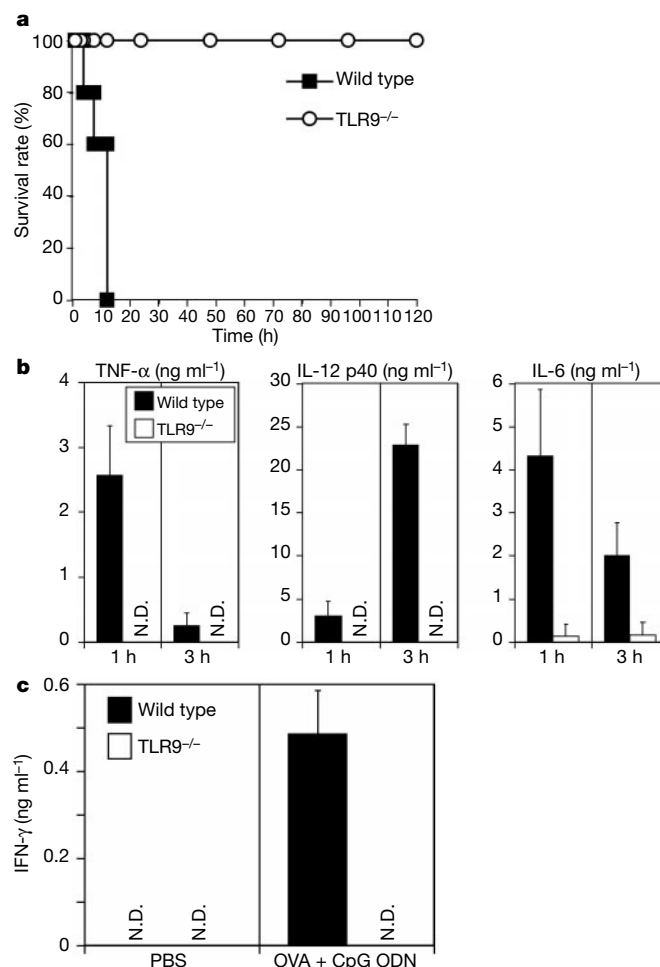


Figure 4 Resistance to CpG ODN-induced shock in TLR9^{-/-} mice. **a**, Age-matched wild-type ($n = 5$) and TLR9^{-/-} mice ($n = 5$) were intraperitoneally injected with CpG ODN (20 nmol) and D-GalN (20 mg). Survival was monitored for 5 d. **b**, Wild-type and TLR9^{-/-} mice were intraperitoneally injected with CpG ODN and D-GalN. Sera were taken at 1 or 3 h after injection. Serum concentrations of TNF-α, IL-12 p40 and IL-6 were measured by ELISA. N.D., not detected. Results are mean of sera samples from three mice. **c**, Age-matched wild-type and TLR9^{-/-} mice were injected with PBS or CpG ODN (5 nmol) plus OVA (150 μg) into the hind footpads. Popliteal lymph node cells were collected after 7 d and cultured with OVA for 24 h. Production of IFN-γ from lymph node cells was analysed by ELISA. Data indicate mean ± s.d. N.D., not detected.

Methods

Cloning of TLR9

A GenBank search resulted in identification of a mouse EST that has a significant similarity with human TLR4. Using PCR-amplified EST as a probe, a full-length cDNA clone containing the complete TLR9 open reading frame was isolated from the mouse RAW 264.7 cDNA library. Human genomic sequence that showed significant homology with the mouse TLR9 gene was found in GenBank. On the basis of this sequence, rapid amplification of cDNA ends protocol was performed to isolate full-length cDNA from U937 cDNA library.

Generation of TLR9^{-/-} mice

The TLR9 genomic DNA was isolated from 129/Sv mouse genomic library and characterized by restriction enzyme mapping and sequencing analysis. The targeting vector was constructed by replacing a 1.0-kb fragment encoding a part of LRR region with a neomycin-resistance gene cassette (*neo*), and a herpes simplex virus thymidine kinase driven by MC1 promoter was inserted into the genomic fragment for negative selection (Fig. 2a). The targeting vector was transfected into embryonic stem cells (E14.1). G418 and gancyclovir doubly resistant colonies were selected and screened by PCR and southern blotting. Homologous recombinants were micro-injected into C57BL/6 blastocysts. Chimaeric mice were mated with C57BL/6 female mice, and heterozygous F₁ progenies were intercrossed in order to obtain TLR9^{-/-} mice. All mice analysed here were F₁ progeny of 129/Ola × C57BL/6.

Reagents

Phosphorothioate-stabilized CpG oligodeoxynucleotide (ODN) (TCC-ATG-ACG-TTC-CTG-ATG-CT)²⁵ was purchased from TIB MOLBIOL or Hokkaido System Science. Phosphorothioate-stabilized non-CpG ODN (GCT-TGA-TGA-CTC-AGC-CGG-AA)⁵ was purchased from Hokkaido System Science. LPS from *Salmonella minnesota* Re-595 and PGN from *Staphylococcus aureus* were purchased from Sigma and Fluka, respectively²¹.

Measurement of cytokine production from macrophages

Thioglycollate-elicited peritoneal macrophages were cultured with the indicated concentrations of CpG ODN, LPS or PGN for 24 h. Concentrations of TNF-α, IL-6 and IL-12 p40 in the culture supernatants were measured by ELISA.

Preparation and analysis of dendritic cells

Bone-marrow cells from wild-type or TLR9^{-/-} mice were cultured with 10 ng ml⁻¹ mouse granulocyte macrophage-colony stimulating factor (Peprotech) in RPMI1640 medium supplemented with 10% fetal bovine serum. At day 6, immature DCs were collected and cultured in the absence or presence of 0.1 μM CpG ODN or 0.1 μg ml⁻¹ LPS in a fresh medium for a further 2 days. The concentration of IL-12 p40 in the culture supernatants was measured by ELISA. DCs were stained with biotinylated antibodies against CD40, CD80, CD86 or MHC class II, and developed with PE-conjugated streptavidin. Flow cytometric analysis was performed using a FACSCalibur with CELLQuest software (Becton Dickinson).

EMSA and *in vitro* kinase assay

Thioglycollate-elicited peritoneal macrophages (1×10^6 cells) from wild-type and TLR9^{-/-} mice were stimulated for the indicated periods and then nuclear proteins were extracted. The extracts were incubated with a specific probe containing NF-κB DNA-binding sites, electrophoresed and visualized by autoradiography.

Thioglycollate-elicited peritoneal macrophages were stimulated with 1.0 μM CpG ODN or 1.0 μg ml⁻¹ LPS for the indicated periods. JNK and IRAK activities in cell lysates were measured by *in vitro* kinase assays as described²³.

Cytokine production of presensitized lymph nodes

Age-matched wild-type and TLR9^{-/-} mice were injected subcutaneously with CpG ODN (5 nmol) plus soluble OVA (150 μg) into the hind footpads. Seven days later, popliteal lymph nodes were collected and cultured with 100 μg ml⁻¹ OVA for 24 h. Concentrations of IFN-γ in the culture supernatants were measured by ELISA.

Received 26 June; accepted 3 October 2000.

- Krieg, A. M. Lymphocyte activation by CpG dinucleotide motifs in prokaryotic DNA. *Trends Microbiol.* **4**, 73–76 (1996).
- Lipford, G. B., Heeg, K. & Wagner, H. Bacterial DNA as immune cell activator. *Trends Microbiol.* **6**, 496–500 (1998).
- Yamamoto, S., Yamamoto, T. & Tokunaga, T. The discovery of immunostimulatory DNA sequence. *Spring. Ser. Immunopathol.* **22**, 11–19 (2000).
- Jakob, T., Walker, P. S., Krieg, A. M., Udey, M. C. & Vogel, J. C. Activation of cutaneous dendritic cells by CpG-containing oligodeoxynucleotides: a role for dendritic cells in the augmentation of Th1 responses by immunostimulatory DNA. *J. Immunol.* **161**, 3042–3049 (1998).
- Sparwasser, T. et al. Bacterial DNA and immunostimulatory CpG oligonucleotides trigger maturation and activation of murine dendritic cells. *Eur. J. Immunol.* **28**, 2045–2054 (1998).
- Hartmann, G., Weiner, G. J. & Krieg, A. M. CpG DNA: a potent signal for growth, activation, and maturation of human dendritic cells. *Proc. Natl Acad. Sci. USA* **96**, 9305–9310 (1999).
- Häcker, H. et al. Cell type-specific activation of mitogen-activated protein kinases by CpG-DNA controls interleukin-12 release from antigen-presenting cells. *EMBO J.* **18**, 6973–6982 (1999).
- Wagner, H. Bacterial CpG DNA activates immune cells to signal infectious danger. *Adv. Immunol.* **73**, 329–368 (1999).
- Klinman, D. M., Verthelyi, D., Takeshita, E. & Ishii, K. J. Immune recognition of foreign DNA: a cure for bioterrorism? *Immunity* **11**, 123–129 (1999).
- Krieg, A. M. The role of CpG motifs in innate immunity. *Curr. Opin. Immunol.* **12**, 35–43 (2000).
- Medzhitov, R. & Janeway, C. A. Jr Innate immunity: the virtues of a nonclonal system of recognition. *Cell* **91**, 295–298 (1997).
- Medzhitov, R., Preston-Hurlburt, P. & Janeway, C. A. Jr A human homologue of the *Drosophila* Toll protein signals activation of adaptive immunity. *Nature* **388**, 394–397 (1997).
- Rock, F. L., Hardiman, G., Timans, J. C., Kastelein, R. A. & Bazan, J. F. A family of human receptors structurally related to *Drosophila* Toll. *Proc. Natl Acad. Sci. USA* **95**, 588–593 (1998).
- Takeuchi, O. et al. TLR6: A novel member of an expanding toll-like receptor family. *Gene* **231**, 59–65 (1999).
- Poltorak, A. et al. Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. *Science* **282**, 2085–2088 (1998).
- Hoshino, K. et al. Toll-like receptor 4 (TLR4)-deficient mice are hyporesponsive to lipopolysaccharide: evidence for TLR4 as the *Lps* gene product. *J. Immunol.* **162**, 3749–3752 (1999).
- Yoshimura, A. et al. Recognition of Gram-positive bacterial cell wall components by the innate immune system occurs via Toll-like receptor 2. *J. Immunol.* **163**, 1–5 (1999).
- Brightbill, H. D. et al. Host defense mechanisms triggered by microbial lipoproteins through toll-like receptors. *Science* **285**, 732–736 (1999).
- Aliprantis, A. O. et al. Cell activation and apoptosis by bacterial lipoproteins through Toll-like receptor-2. *Science* **285**, 736–739 (1999).
- Underhill, D. M. et al. The Toll-like receptor 2 is recruited to macrophage phagosomes and discriminates between pathogens. *Nature* **401**, 811–815 (1999).

21. Takeuchi, O. *et al.* Differential roles of TLR2 and TLR4 in recognition of gram-negative and gram-positive bacterial cell wall components. *Immunity* **11**, 443–451 (1999).
22. Takeuchi, O. *et al.* Preferentially the R-stereoisomer of the mycoplasma lipopeptide macrophage-activating lipopeptide-2 activates immune cells through a toll-like receptor 2- and MyD88-dependent signaling pathway. *J. Immunol.* **164**, 554–557 (2000).
23. Kawai, T., Adachi, O., Ogawa, T., Takeda, K. & Akira, S. Unresponsiveness of MyD88-deficient mice to endotoxin. *Immunity* **11**, 115–122 (1999).
24. Häcker, H. *et al.* Immune cell activation by bacterial CpG-DNA through myeloid differential marker 88 and tumor necrosis factor receptor-associated factor (TRAF)6. *J. Exp. Med.* **192**, 595–600 (2000).
25. Sparwasser, T. *et al.* Macrophages sense pathogens via DNA motifs: induction of tumor necrosis factor- α -mediated shock. *Eur. J. Immunol.* **27**, 1671–1679 (1997).
26. Lipford, G. B. *et al.* CpG-DNA-mediated transient lymphadenopathy is associated with a state of Th1 predisposition to antigen-driven responses. *J. Immunol.* **165**, 1228–1235 (2000).
27. Liang, H., Reich, C. F., Pisetsky, D. S., Lipsky, P. E. The role of cell surface receptors in the activation of human B cells by phosphorothioate oligonucleotides. *J. Immunol.* **165**, 1438–1445 (2000).
28. Krieg, A. M. *et al.* CpG motifs in bacterial DNA trigger direct B-cell activation. *Nature* **374**, 546–549 (1995).
29. Macfarlane, D. E. & Manzel, L. Antagonism of immunostimulatory CpG-oligodeoxynucleotides by quinacrine, chloroquine, and structurally related compounds. *J. Immunol.* **160**, 1122–1131 (1998).
30. Häcker, H. *et al.* CpG-DNA-specific activation of antigen-presenting cells requires stress kinase activity and is preceded by non-specific endocytosis and endosomal maturation. *EMBO J.* **17**, 6230–6240 (1998).
31. Yi, A. K. & Krieg, A. M. Rapid induction of mitogen-activated protein kinases by immune stimulatory CpG DNA. *J. Immunol.* **161**, 4493–4497 (1998).

Acknowledgements

We thank G. B. Lipford for helpful discussions; T. Aoki for secretarial assistance; and N. Tsuji, N. Iwami and E. Nakatani for technical assistance. We also thank Hayashibara Biochemical Laboratories, Inc. for providing anti-IRAK antibody. This work was supported in part by grants from the Ministry of Education, Science, Sports and Culture of Japan, and Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

Correspondence and requests for materials should be addressed to S.A. (e-mail: sakira@biken.osaka-u.ac.jp). Sequences have been deposited in GenBank under accession numbers AB045180 and AB045181 for human and mouse TLR9, respectively.

Structure of the bacteriophage ϕ 29 DNA packaging motor

Alan A. Simpson*†, Yizhi Tao*†‡, Petr G. Leiman*,
Mohammed O. Badasso§, Yongning He*, Paul J. Jardine||,
Norman H. Olson*, Marc C. Morais*, Shelley Grimes§,
Dwight L. Anderson§, Timothy S. Baker* & Michael G. Rossmann*

* Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907-1392, USA

§ Departments of Microbiology and Oral Science, 18-246 Moos Tower, University of Minnesota, Minneapolis, Minnesota 55455, USA

|| Department of Biology, University of New Brunswick, Fredericton, New Brunswick E3B 6E1, Canada

† These authors contributed equally to this work

Motors generating mechanical force, powered by the hydrolysis of ATP, translocate double-stranded DNA into preformed capsids (proheads) of bacterial viruses^{1,2} and certain animal viruses³. Here we describe the motor that packages the double-stranded DNA of the *Bacillus subtilis* bacteriophage ϕ 29 into a precursor capsid. We determined the structure of the head–tail connector—the central component of the ϕ 29 DNA packaging motor—to 3.2 Å resolution by means of X-ray crystallography. We then fitted the connector into the electron densities of the prohead and of the partially packaged prohead as determined using cryo-electron microscopy and image reconstruction analysis. Our results suggest that the prohead plus dodecameric connector, prohead RNA, viral ATPase and DNA comprise a rotary motor with the

head–prohead RNA–ATPase complex acting as a stator, the DNA acting as a spindle, and the connector as a ball-race. The helical nature of the DNA converts the rotary action of the connector into translation of the DNA.

The bacteriophage ϕ 29 (Fig. 1) is a 19-kilobase (19-kb) double-stranded DNA (dsDNA) virus with a prolate head and complex structure⁴. The prohead (Fig. 1), into which the DNA is packaged, is about 540 Å long and 450 Å wide⁵. The ϕ 29 connector, a cone-shaped dodecamer of gene product 10 (gp10), occupies the pentagonal vertex at the base of the prohead⁵ and is the portal for DNA entry during packaging and DNA ejection during infection⁶. The connector, in association with the oligomeric, ϕ 29-encoded prohead RNA (pRNA) and a viral ATPase (gp16), is required for DNA packaging^{7–9}. However, only the first 120 bases of the 174-base pRNA are essential for packaging⁷. The covalent adduct of the genomic dsDNA with gp3 (DNA–gp3) can be packaged into proheads in about three minutes *in vitro* (P.J.J., unpublished results). The connector proteins of tailed phages⁶ vary in relative molecular mass (M_r) from 36,000 (36K) in ϕ 29 to 83K in phage P22, and assemble into oligomers with a central channel. The structure of the isolated ϕ 29 connector has been studied by atomic force microscopy¹⁰ and cryo-electron microscopy (cryo-EM) of two-dimensional arrays¹¹, immuno-electron microscopy¹² and X-ray crystallography^{13,14}.

The connector structure, as now determined by X-ray crystallography, can be divided into three, approximately cylindrical regions: the narrow end, the central part, and the wide end, having external radii (Å) of 33, 47 and 69, respectively (Fig. 2). These regions are respectively 25, 28 and 22 Å in height, making the total connector 75 Å long. The internal channel has a diameter of about 36 Å at the narrow end, increasing to 60 Å at the wide end. Comparison with electron microscopy reconstructions^{5,11} shows that the narrow end protrudes from the portal vertex of the phage head, is associated with the multimeric pRNA, and binds the lower collar in the mature virus.

The electron density of the connector was interpreted in terms of the amino-acid sequence¹⁵ and was confirmed by the two Hg sites (see Methods section) corresponding to the only cysteine residues in the sequence. Residues 1 to 11, 229 to 246, and 287 to 309 at the carboxy terminus are not seen in the electron density. The second and third disordered regions are both located on the inside of the channel, close to the junction of the central and wide regions. The structure is dominated by three long helices (α 1, α 3 and α 5) in each monomer that run the length of the central region, joining the two end domains of the connector (Fig. 2). These helices are arranged at an angle of about 40° with respect to the central 12-fold axis. The two end domains are composed predominantly of β -sheets and extended polypeptide chains. Immuno-electron microscopy mapping of the sequence onto the connector surface¹² is consistent with the X-ray structure only as far as localization of the external amino-terminal residues with the pRNA-binding region at the narrow end of the connector. The RNA recognition motif structure, previously predicted for the N-terminal regions of the connector monomer^{16,17}, is not present in the structure.

The surface of the monomer presents a net negative charge to one neighbour and a net positive charge to the other neighbour, possibly aiding the assembly of the dodecamer. The exterior of the connector has no significant regions of charge accumulation, implying that its rotation might be facilitated by its oily, smooth, external surface. However, the basic character of the disordered 11 amino-terminal residues could alter the surface properties to some extent and may facilitate interaction with the pRNA. In contrast, the inside of the channel has a preponderance of negative charge at its wide end, which may repel the DNA, permitting its smooth passage during packaging and ejection. The channel through which messenger RNA is translocated in reoviruses has similar properties¹⁸.

We have determined the structures of four distinct types of ϕ 29

† Present address: Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA.

particles by cryo-EM and image reconstruction analyses (Fig. 1, Table 1): the $\phi 29$ prohead including pRNA, the prohead after removal of the pRNA, the partially DNA-packaged prohead including the pRNA and ATPase (gp16), and (data not shown) empty proheads found in the *in vitro* packaging system. A difference map between proheads with and without the pRNA locates the pRNA at the narrow end of the connector, in agreement with ref. 19. The difference map (Fig. 1d) was based on reconstructions in which the five-fold symmetry had been imposed along the long axis of the particles. If the pRNA were six-fold symmetric as suggested by genetic data^{8,9}, the resultant averaged density would be weak and smeared. However, we observe that the pRNA has excellent density, higher by a factor of ~ 1.5 than the density of the head, thus demonstrating that the pRNA is consistent with the imposed five-fold symmetry. Furthermore, a reconstruction computed without imposing five-fold symmetry (Table 1, data set b'; M.C.M., Y.T., P.J.J., D.L.A., T.S.B. and M.G.R., manuscript in preparation; also see the Methods section) also showed good pRNA density with almost exact five-fold symmetry.

The difference density representing the pRNA consists of a continuous ring with internal and external radii of 34 and 62 Å, respectively. In addition, there are five spokes radiating slightly outwards, which are in the same orientations relative to the head in all reconstructions (Fig. 1). The alignment of the pRNA spokes with the prohead implies that either the pRNA communicates with the prohead through the connector or that the pRNA and the head directly interact. Indeed, the cryo-EM reconstructions of the prohead show density connecting the pRNA and head at a contour level well above that of noise in the background (Fig. 3a). Although genetic studies^{8,9} had shown that the pRNA forms hexamers by intermolecular base pairing, only a pentameric model could be readily fitted into the cryo-EM density, with the five A helices fitting into the radial spokes (Fig. 3b). A hexameric ring would have had too large a radius. It is possible that a hexameric pRNA might be required to initiate pRNA binding, but that at some stage one pRNA molecule^{8,9} is shifted out of the ring or eliminated. As the handedness of the pRNA density is not known, the direction of the pRNA helices within the ring is uncertain.

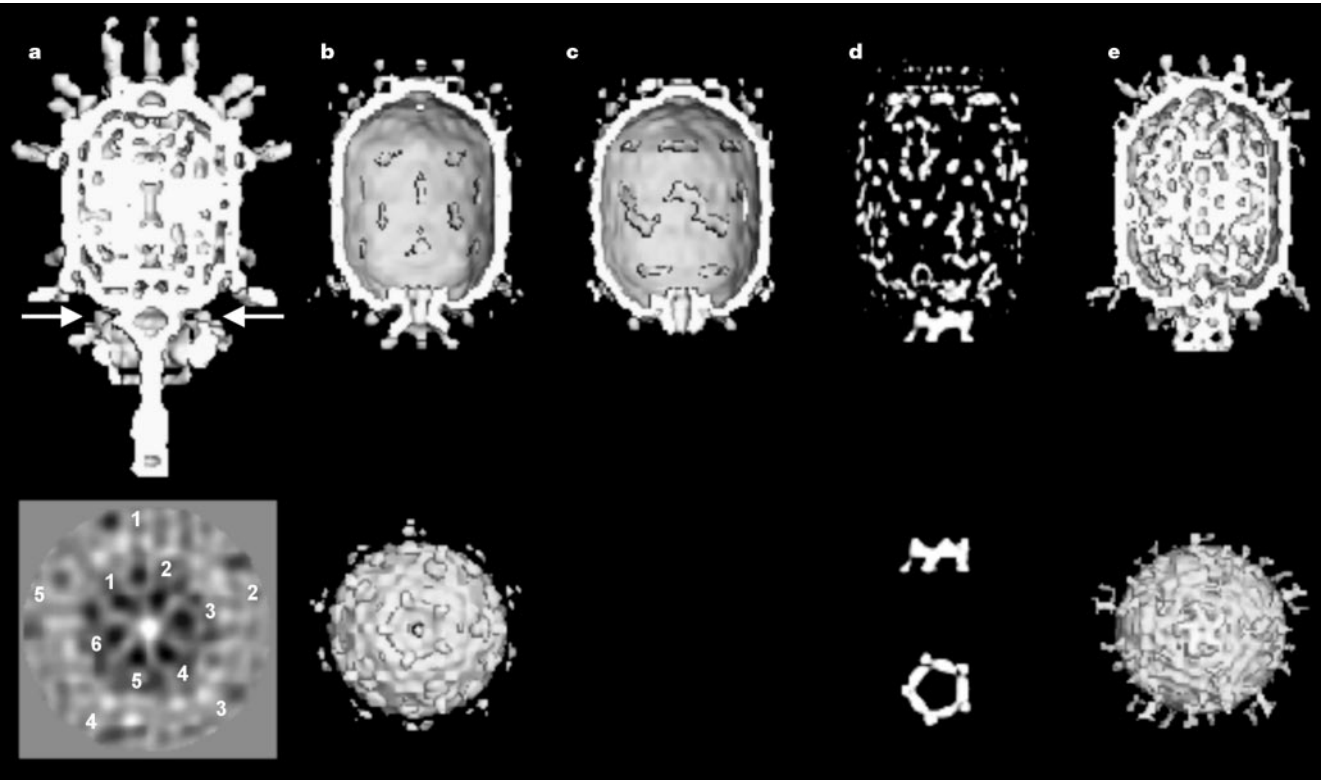


Figure 1 Cryo-EM reconstructions. Top and bottom rows: **a**, mature $\phi 29$; **b**, prohead + 120-base pRNA; **c**, prohead treated with RNase; **d**, difference map between **b** and **c**; and **e**, partially packaged particle (with DNA in channel). End-on views, looking along the tail towards the head, are given for **b** and **e**. Orthogonal difference pRNA densities are shown

below the difference map in **d**. The arrow in **a** shows the position of the section below that was obtained by averaging particles without imposing five-fold symmetry. Note that both the five-fold symmetry of the downward-pointing head fibres (numbered 1 to 5) and six-fold symmetry (numbered 1 to 5) of the lower collar are visible.

Table 1 Cryo-EM data and image reconstruction

Data set*	Prohead+pRNA		RNase-treated prohead	Prohead+pRNA+DNA-gp3+gp16
	(b)	(b')	(c)	(e)
Symmetry imposed	5	1	5	5
Underfocus (μm)†	2.3	2.3	2.3	2.3
Number of particles‡	609 (1,153)	680 (1,153)	285 (4,81)	534 (1,038)
Correlation coefficient§	0.324(0.047)	0.477 (0.061)	0.263 (0.056)	0.236 (0.050)
Resolution (Å)	26	35	35	32

* The data set identifications b, c and e correspond to the reconstructions shown in Fig. 1 (top and bottom). The reconstructions in column b' (not shown in Fig. 1) uses the same data as in column b, but with a lower imposed symmetry.
† Determined from the contrast transfer function of the microscope.
‡ The number of particles included in each three-dimensional map. The total number of boxed particles is given in parentheses.
§ Real-space correlation coefficients averaged over all particles and standard deviation (in parentheses).

The X-ray structure of the connector fits accurately into the portal vertex of the prohead reconstruction (Fig. 3a). The wide end of the connector contacts the inside of the head, whereas the narrow end protrudes from the portal vertex where it is encircled by the pentameric pRNA, which makes five, symmetrically disposed, interactions with the head (Fig. 3a). It appears that the orientation of the pRNA is determined by its interaction with the head rather than with the connector. Although the contact region of pRNA and connector is substantial, the oily surface of the connector and its 5–12 symmetry mismatch would aid its smooth rotation, as proposed in ref. 1. Residues 229 to 246 and 287 to 309, disordered in the crystal structure, occur in the DNA channel, roughly at the border between the central and wide regions of the connector, which corresponds to a low-density extension of the DNA channel (Fig. 3a). The partially disordered residues include the unusual

amino-acid sequences Glu-Lys-Lys-Glu-Arg and Arg-Arg-Glu, which might be ideal for interacting with the charged sugar-phosphate DNA backbone.

Actively packaging prohead samples were flash-frozen 2.5 min after initiation of packaging. The resulting cryo-EM images were divided, by visual inspection, into two groups, one representing particles that appeared to be empty (data not shown) and the other those that were partially filled (Fig. 1e; Table 1, data set e). The reconstruction of the empty particles was similar to that of the prohead including the pRNA (Fig. 1b; Table 1, data set b), whereas the partially filled prohead reconstruction showed both the five-fold pRNA structure and additional density associated with each of the five rotary spokes (Fig. 1e), implying that the components corresponding to this density are involved in DNA packaging. This density, therefore, might be attributed to the viral ATPase (gp16).

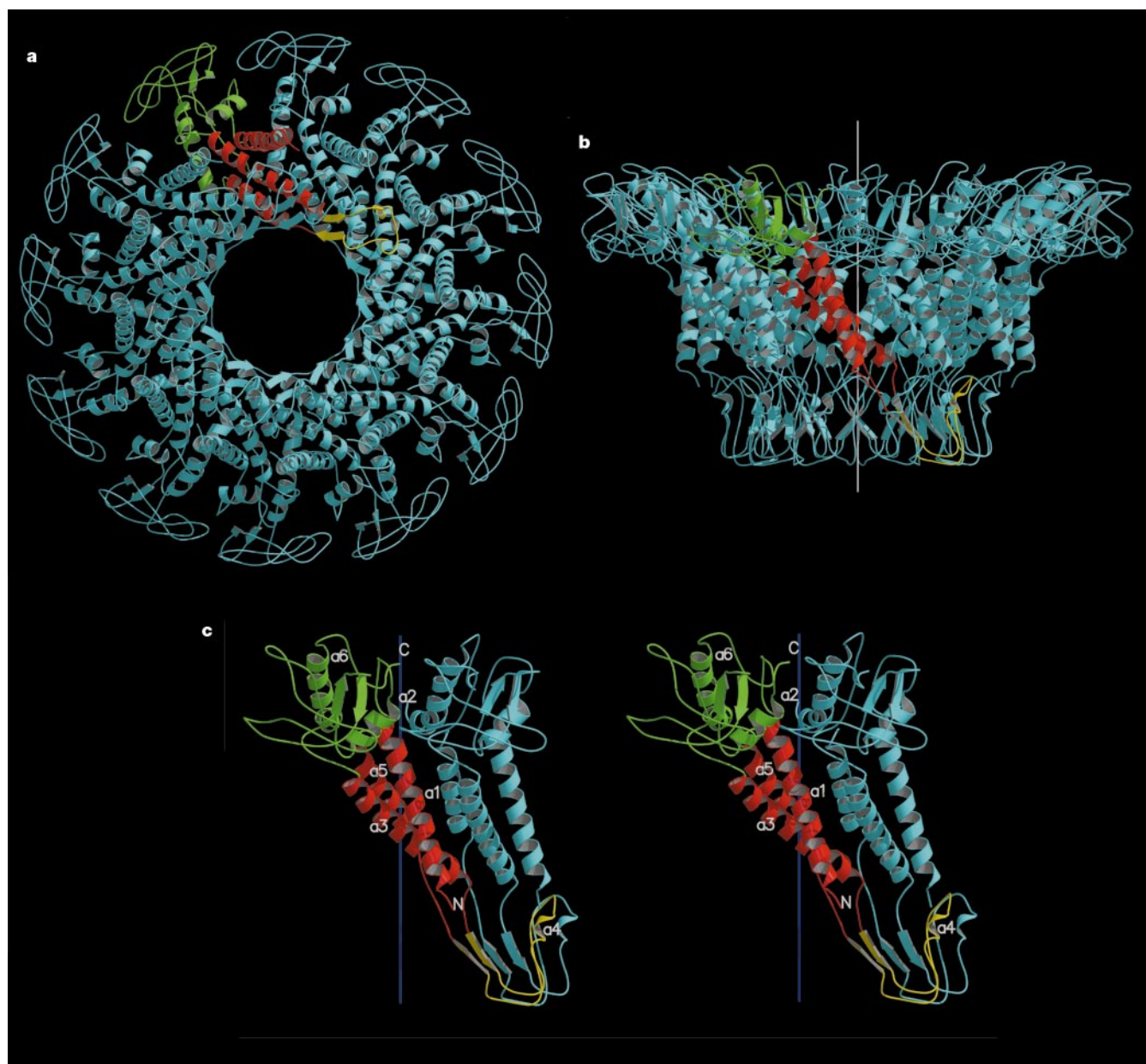


Figure 2 Connector structure ribbon diagrams. **a**, The dodecameric connector seen from the tail looking towards the head; **b**, a side view with the pRNA-binding site at the bottom, showing the conical structure of the connector and the helical twist of each subunit around the 12-fold axis (white); **c**, a stereo diagram of a pair of monomers. One monomer is coloured red in the central domain, green in the wide-end domain that resides inside the capsid, and yellow at the narrow-end domain. The other monomers are all coloured blue.

The ordered part of the polypeptide starts with helix $\alpha 1$ on the outside of the connector, going towards the wide end (residues 61 to 128 and 247 to 286). Helix $\alpha 3$ (residues 129 to 157) returns the chain to the narrow end (residues 158 to 202). The tip of the connector at the narrow end is formed by residues 164 to 170 and 185 to 196. Helix $\alpha 5$ (residues 208 to 226) returns the polypeptide to the wide end through the second disordered section. (Drawn with the program MOLSCRIPT²⁵ and RASTER3D²⁶.)

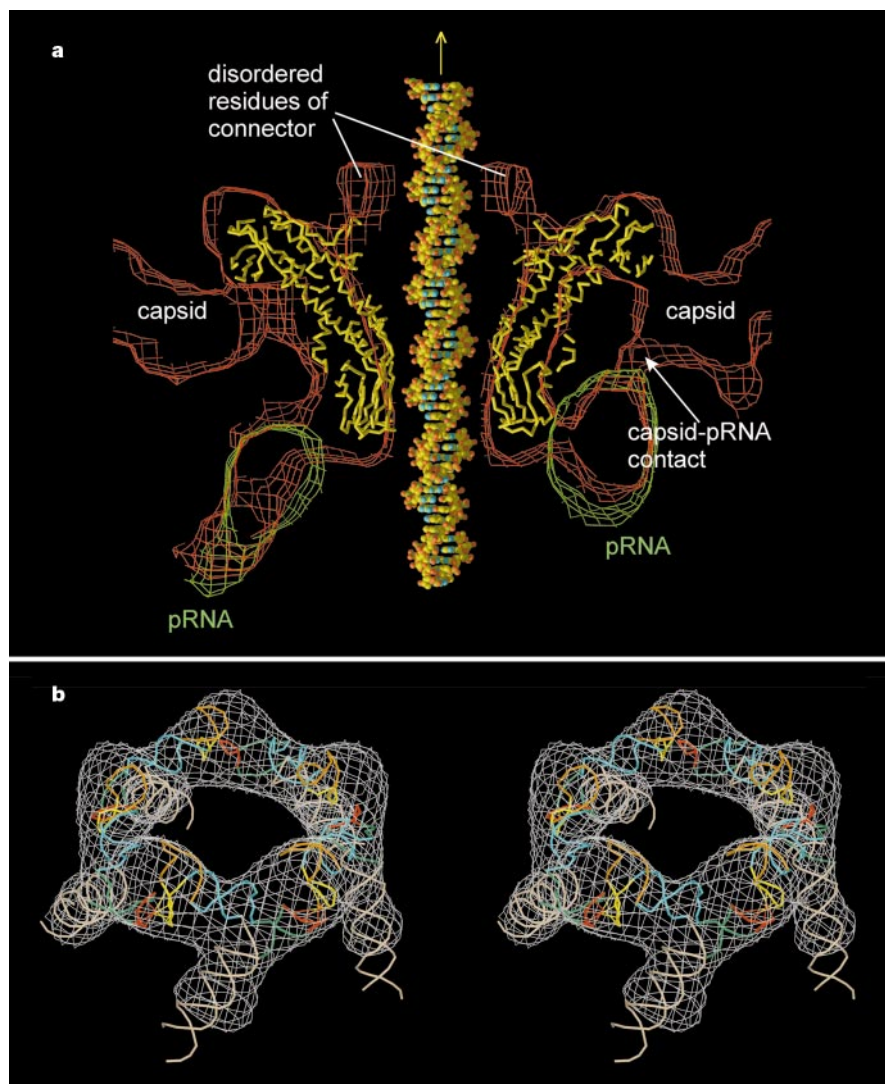


Figure 3 Cryo-EM density fitted with atomic structures. **a**, Cross-section of the cryo-EM prohead density (red) fitted with the C α backbone of the connector (yellow) and the cryo-EM pRNA difference map (green). Shown also is a DNA molecule placed through the central channel of the connector. The prohead capsid, one of the five contacts between the pRNA with the capsid, and the partially disordered residues 229 to 246 and 287 to 309 in the connector are indicated. **b**, Stereo drawing of the 120-base pRNA pentamer

fitted into the cryo-EM density. Secondary structural elements²⁷ are shown in white (A helix), blue (C helix), orange (E helix), yellow (CE loop), green (D helix) and red (D loop). Four intermolecular base pairs are shown in yellow-red. It had previously been shown that helices D, C and E participate in binding pRNA to the connector, and it was suggested that helix A binds the viral ATPase (gp16)²⁸. (Drawn with the programs XTALVIEW²⁵ and RASTER3D²⁵.)

The partially filled particles contain less DNA in the head as compared to mature virions, indicating that the packaging process was incomplete at the time of freezing. This cryo-EM image reconstruction also shows some density inside the narrow end of the connector channel (Fig. 1e), which can be interpreted as DNA entering the prohead.

The accumulated data provide the basis for proposing different mechanisms for the DNA packaging machine. Here we discuss one proposal that is consistent with the structural observations (Fig. 4). The DNA, connector and prohead–pRNA–ATPase complex form a set of concentric structures with 10₁-, 12- and 5-fold symmetry, respectively. These constitute a movable central spindle, intervening ball-race, and a static outer assembly (stator), which powers the motor. The sequential firing of the ATPases (I to V) powers a series of conformational changes that translate the DNA into the head (Fig. 4). As successive ATPases occur one-fifth of a rotation around the circumference of the DNA, the DNA must either rotate by one-fifth of a turn or translate by one-fifth of its pitch to maintain the same relationship to the currently active ATPase (compare Fig. 4a and c). The large size of the DNA would make translation energetically

favourable. By the time each of the five ATPases have fired, the connector will have rotated 60°, and the DNA will have been translated the length of one pitch of its helix, thus packaging 10 base pairs. This corresponds to a movement of two base pairs per ATP hydrolysis, consistent with the observed ATP consumption during packaging²⁰.

The narrow- and wide-end domains of each connector monomer differ in their orientation by up to 2° relative to the central helices, which are twisted around the central axis and are roughly orthogonal to the DNA backbone. The flexibility of the end domains would allow the connector to rotate by a two-step mechanism (Fig. 4). Thus, the connector may have a more complex role than a ball-race because it may also be required to undergo a helical spring-like oscillation that couples the rotational motor action to the linear translocation of the DNA spindle.

The DNA packaging mechanism utilizes the symmetry mismatch and ATP hydrolysis to effect connector rotation as proposed in refs 1 and 21. However, the connector does not rotate along the DNA helix, as in a rigid nut–bolt assembly¹; rather, each connector monomer occupies a different position on the DNA helix after

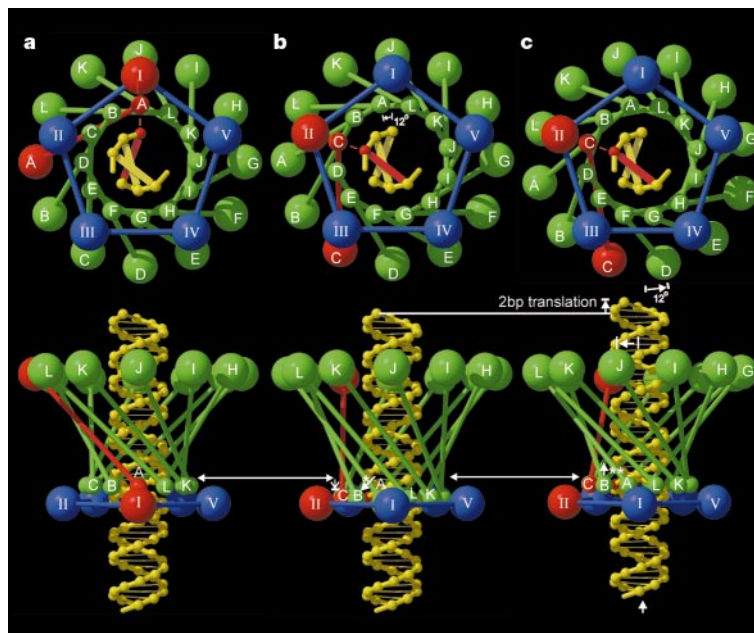


Figure 4 The DNA packaging mechanism. Shown is one cycle in the mechanism that rotates the connector and translates the DNA into the head. The view down the connector axis (top) is towards the head, whereas the bottom row shows side views corresponding to that seen in Fig. 2b. Eleven of the 12 subunits (A, B, ..., L) of the connector are shown in green; the 'active' monomer is shown in red. The connector is represented as a set of small spheres at the narrow end and a set of larger spheres at the wide end connected by a line representing the central helical region. The pRNA-ATPase complexes (I, II, III, IV, V), surrounding the narrow end, are shown by a set of four blue spheres and one red sphere. The DNA base aligned with the active connector monomer is also shown in red. In **a**, the

active pRNA-ATPase I interacts with the adjacent connector monomer (A), which in turn contacts the aligned DNA base. In **b**, the narrow end of the connector has moved anti-clockwise by 12° to place the narrow end of monomer C opposite ATPase II, the next ATPase to be fired, causing the connector to expand lengthwise by slightly changing the angle of the helices in the central domain (white arrow with asterisk). In **c**, the wide end of the connector has followed the narrow end, while the connector relaxes and contracts (white arrow with two asterisks), thus causing the DNA to be translated into the phage head. For the next cycle, ATPase II is activated, causing the connector to be rotated another 12°, and so forth. (Drawn with the program RASTER3D²⁵.)

every ATP hydrolysis, as proposed in ref. 21. The energy for DNA packaging is supplied by the external stator composed of the head, pRNA and ATPase.

There are several well characterized examples of molecular motors. These fall into two classes, the first of which consists of rotary motors, such as the F₁ ATPase²² and the bacterial flagellar motor²³, which turn a spindle by a rotary conformational change mechanism. The second class consists of linear motors, like those required for muscle contraction²⁴, where conformational fluctuations undergo a translation with each step. The ϕ 29 DNA packaging motor as we have described it represents a third class, where rotary motion is converted into translational movement. □

Methods

X-ray crystallography

Bacteriophage ϕ 29 connectors were produced in *Escherichia coli* and were crystallized¹⁴ into several different space groups. The crystals used for the present studies belonged to space group C2, with $a = 177.2$, $b = 169.2$, $c = 185.4$ Å, $\beta = 114.10^\circ$, consistent with one connector per crystallographic asymmetric unit. X-ray diffraction data were collected at the Cornell High Energy Synchrotron Source (CHESS), the Advanced Photon Source (APS) BioCARS beamline 14BMc, and the APS Structural Biology Center beamline 19ID. The data used had an R_{merge} of 6.5% to 3.2 Å resolution (24.0% in the highest resolution shell). The Hg salt, thimerosal, was used as a heavy atom derivative. The diffraction data of this derivative had an R_{merge} of 10.8% (26.0%) to 3.4 Å resolution and a difference in structure amplitudes of 24% with respect to the native data.

A self-rotation function showed that the connector had 12-fold symmetry with the axis perpendicular to the unique monoclinic b axis, and a Patterson self-translation function²⁵ showed that the axis passed through one of the crystallographic dyads. A three-dimensional, low-resolution, electron density distribution was constructed of the connector, based on orthogonal views of the connector derived from cryo-EM images of two-dimensional, tilted arrays¹¹. The position along and rotation around the 12-fold axis of the model was determined by structure factor correlation searches using the program AMoRe²⁵ and packing considerations. The initial model phases were extended from 20 Å to 3.5 Å resolution (A.A.S. *et al.*, manuscript in preparation) by 12-fold, non-crystallographic symmetry (NCS), electron density averaging²⁵ and solvent flattening using the program DM²⁵. The resultant map was not good enough for interpretation, but

did permit the solution of the Hg sites in a 4.0 Å resolution difference map. Two independent sites per monomer were found, each with 12 sites related by the NCS. The sites were refined with the program MLPHARE and used to compute phases to 3.2 Å resolution, which were then improved by iterative density NCS averaging and solvent flattening. The final map could be readily interpreted in terms of the amino-acid sequence¹⁵ using the program O²⁵.

Cryo-electron microscopy

The preparation of specimens for cryo-EM was performed as described previously⁵. The model-based polar-Fourier-transform method²⁶ was used to determine the orientation of each particle. To confirm the five-fold symmetry of the pRNA, the five-fold symmetry imposed on the initial reconstruction was relaxed. Although the orientation of the particles in each image had been determined with respect to the five-fold symmetric head, a subsequent search was performed over the five different possible orientations, and the images were averaged without imposing five-fold symmetry (M.C.M., Y.T., P.J.J., D.L.A., T.S.B. and M.G.R., manuscript in preparation). This procedure was applied to images of proheads and native virus. A section through the reconstruction, perpendicular to the phage axis (Fig. 1), shows both the five-fold symmetry of the head fibres and the six-fold symmetry of the lower collar in the same section, demonstrating that the procedure did not unintentionally impose five-fold symmetry.

In their cryo-EM image reconstruction of ϕ 29 proheads, Ibarra *et al.*¹⁹ attributed the five, roughly radial, pRNA spokes protruding from the connector to accidental inclusion of proheads in an inverted orientation, causing superposition of the head fibres onto the pRNA. However, in our samples all particles were lying almost flat on the grid, which made it easy to distinguish the two ends of the head. Furthermore, the presumed pRNA features are of much higher density than the head fibres, showing that the head fibres could not give rise to the RNA spokes. Misoriented particles also should have produced an image of the connector at both ends of the head and would have resulted in 'fibres' on the connector in the reconstruction of the RNase-treated prohead that is missing the pRNA. None of these features was observed.

Received 24 July; accepted 22 September 2000.

- Hendrix, R. W. Symmetry mismatch and DNA packaging in large bacteriophages. *Proc. Natl Acad. Sci. USA* **75**, 4779–4783 (1978).
- Hendrix, R. W. Bacteriophage DNA packaging: RNA gears in a DNA transport machine. *Cell* **94**, 147–150 (1998).
- Zhou, Z. H., Chen, D. H., Jakana, J., Rixon, F. J. & Chiu, W. Visualization of tegument-capsid interactions and DNA in intact herpes simplex virus type 1 virions. *J. Virol.* **73**, 3210–3218 (1999).
- Anderson, D. L., Hickman, D. D. & Reilly, B. E. Structure of *Bacillus subtilis* bacteriophage ϕ 29 and the length of ϕ 29 deoxyribonucleic acid. *J. Bacteriol.* **91**, 2081–2089 (1966).

5. Tao, Y. *et al.* Assembly of a tailed bacterial virus and its genome release studied in three dimensions. *Cell* **95**, 431–437 (1998).
6. Valpuesta, J. M. & Carrascosa, J. L. Structure of viral connectors and their function in bacteriophage assembly and DNA packaging. *Q. Rev. Biophys.* **27**, 107–155 (1994).
7. Guo, P., Erickson, S. & Anderson, D. A small viral RNA is required for in vitro packaging of bacteriophage ϕ 29 DNA. *Science* **236**, 690–694 (1987).
8. Zhang, F. *et al.* Function of hexameric RNA in packaging of bacteriophage ϕ 29 DNA *in vitro*. *Mol. Cell* **2**, 141–147 (1998).
9. Guo, P., Zhang, C., Chen, C., Garver, K. & Trotter, M. Inter-RNA interaction of phage ϕ 29 pRNA to form a hexameric complex for viral DNA transportation. *Mol. Cell* **2**, 149–155 (1998).
10. Müller, D. J., Engel, A., Carrascosa, J. L. & Vézé, M. The bacteriophage ϕ 29 head-tail connector imaged at high resolution with the atomic force microscope in buffer solution. *EMBO J.* **16**, 2547–2553 (1997).
11. Valpuesta, J. M., Fernández, J. J., Carazo, J. M. & Carrascosa, J. L. The three-dimensional structure of a DNA translocating machine at 10 Å resolution. *Structure* **7**, 289–296 (1999).
12. Valle, M. *et al.* Domain architecture of the bacteriophage ϕ 29 connector protein. *J. Mol. Biol.* **288**, 899–909 (1999).
13. Guasch, A. *et al.* Crystallographic analysis reveals the 12-fold symmetry of the bacteriophage ϕ 29 connector particle. *J. Mol. Biol.* **281**, 219–225 (1998).
14. Badasso, M. O. *et al.* Purification, crystallization, and initial X-ray analysis of the head-tail connector of bacteriophage ϕ 29. *Acta Crystallogr. D* **56**, 1187–1190 (2000).
15. Vleck, C. & Paces, V. Nucleotide sequence of the late region of *Bacillus* ϕ 29 completes the 19285-bp sequence of ϕ 29 genome. Comparison with the homologous sequence of phage PZA. *Gene* **46**, 215–225 (1986).
16. Grimes, S. & Anderson, D. RNA dependence of the bacteriophage ϕ 29 DNA packaging ATPase. *J. Mol. Biol.* **215**, 559–566 (1990).
17. Donate, L. E., Valpuesta, J. M., Mier, C., Rojo, F. & Carrascosa, J. L. Characterization of an RNA-binding domain in the bacteriophage ϕ 29 connector. *J. Biol. Chem.* **268**, 20198–20204 (1993).
18. Reinisch, K. M., Nibert, M. L. & Harrison, S. C. Structure of the reovirus core at 3.6 Å resolution. *Nature* **404**, 960–967 (2000).
19. Ibarra, B. *et al.* Topology of the components of the DNA packaging machinery in the phage ϕ 29 prohead. *J. Mol. Biol.* **298**, 807–815 (2000).
20. Guo, P., Peterson, C. & Anderson, D. Prohead and DNA-gp3-dependent ATPase activity of the DNA packaging protein gp16 of bacteriophage ϕ 29. *J. Mol. Biol.* **197**, 229–236 (1987).
21. Dube, P., Tavares, P., Lurz, R. & van Heel, M. The portal protein of bacteriophage SPP1: a DNA pump with 13-fold symmetry. *EMBO J.* **12**, 1303–1309 (1993).
22. Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. Structure at 2.8 Å resolution of F_1 -ATPase from bovine heart mitochondria. *Nature* **370**, 621–628 (1994).
23. Silverman, M. R. & Simon, M. I. Flagellar rotation and the mechanism of bacterial motility. *Nature* **249**, 73–74 (1974).
24. Vale, R. D. & Milligan, R. A. The way things move: looking under the hood of molecular motor proteins. *Science* **288**, 88–95 (2000).
25. Rossmann, M. G. & Arnold, E. *International Table for Crystallography* Vol. F (International Union of Crystallography, Kluwer Academic, Dordrecht) (in the press).
26. Baker, T. S. & Cheng, R. H. A model-based approach for determining orientations of biological macromolecules imaged by cryoelectron microscopy. *J. Struct. Biol.* **116**, 120–130 (1996).
27. Bailey, S. *et al.* Phylogenetic analysis and secondary structure of the *Bacillus subtilis* bacteriophage RNA required for DNA packaging. *J. Biol. Chem.* **265**, 22365–22370 (1990).
28. Reid, R. J. D., Bodley, J. W. & Anderson, D. Identification of bacteriophage ϕ 29 prohead RNA domains necessary for *in vitro* DNA-gp3 packaging. *J. Biol. Chem.* **269**, 9084–9089 (1994).

Acknowledgements

We thank J. Carrascosa for comments and discussions of this paper. We also thank the staff of CHESS, APS BioCARS and APS Structural Biology Center for assistance in data collection. This work was supported by the NSF (M.G.R.), the NIH (D.L.A. and T.S.B.), and an NSF Shared Instrumentation grant (T.S.B. and M.G.R.).

Correspondence and requests for materials should be addressed to M.G.R. (e-mail: mgr@indiana.bio.purdue.edu). Coordinates of the connector (accession number 1FOU) and pRNA (accession number 1FOQ) have been deposited with the Protein Data Bank.

errata

Logical computation using algorithmic self-assembly of DNA triple-crossover molecules

Chengde Mao, Thomas H. LaBean, John H. Reif & Nadrian Seeman

Nature **407**, 493–496 (2000).

In Fig. 2 of this paper, y_2 should have equalled 0 in Calculation 1; and x_2 , y_3 and y_4 should have equalled 0 in Calculation 2. □

Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model

Peter M. Cox, Richard A. Betts, Chris D. Jones, Steven A. Spall & Ian J. Totterdell

Nature **408**, 184–187 (2000).

In this paper, part of the labelling of Figs 2–4 was printed incorrectly. In Fig. 2, the year 2950 should have been 2050. In Figs 2 and 3, the year 1050 should have been 1950. In Fig. 4, the numbers on the y axis should have been labelled sequentially from –200 to +400; in addition, in Fig. 4a the y axis should have been labelled ‘Changes in vegetation carbon (Gt C)’. □

correction

Allometric scaling of production and life-history variation in vascular plants

Brian J. Enquist, Geoffrey B. West, Eric L. Charnov & James H. Brown

Nature **401**, 907–911 (1999).

In the last sentence of the abstract the relative growth rate equation, presented as $(1/M)(dM/df)$, should read $(1/M)(dM/dt)$. Also, the sentence following that containing equation (5) should read “Thus, regardless of any possible time dependence of either the proportionality constants or the density, a plot of $M^{1/4}$ versus $M_0^{1/4}$ for fixed times t and t_0 for any species should yield a straight line with a universal slope of unity but with an intercept that depends on the time interval and the species.”

In other words, the primary relationship is in terms of mass, M and not diameter, D . Equations (4) and (5), which relate the dependence of trunk diameter, D , on time, t , are only valid if the proportionality constant $A \equiv C_D/\rho^{3/8}$ is time-independent; A is the proportionality constant in the allometric relation $D = AM^{3/8}$ of equation (3). This is needed because data are typically given in terms of D rather than M . The other proportionality constants, C_G and C_B , occurring, respectively, in the growth equation (2) and allometric equation for metabolic rate (3), can be time-dependent.

As our analysis of the data assumed time-independence of all of these coefficients as well as of wood density, ρ (see below equation (6)), this oversight does not affect our results or conclusions. If A were slowly varying, it would contribute a small correction to the unit slope prediction given by $(t - t_0)d$, where d is the logarithmic derivative of A . This is expected to be very small, especially as the time interval $(t - t_0)$ is small relative to the lifespan of the species sampled. Furthermore, any time dependence in A is almost certainly smaller than its variation across species and effects arising from neglecting maintenance costs in the growth equation (2). We thank J. Banavar, J. Damuth, A. Maritan and A. Rinaldo for bringing this oversight to our attention. □

Life and death matters

Looking at living cells, and looking at cell death, in a cell biology selection.

Axiovert 200 with Cell Observer

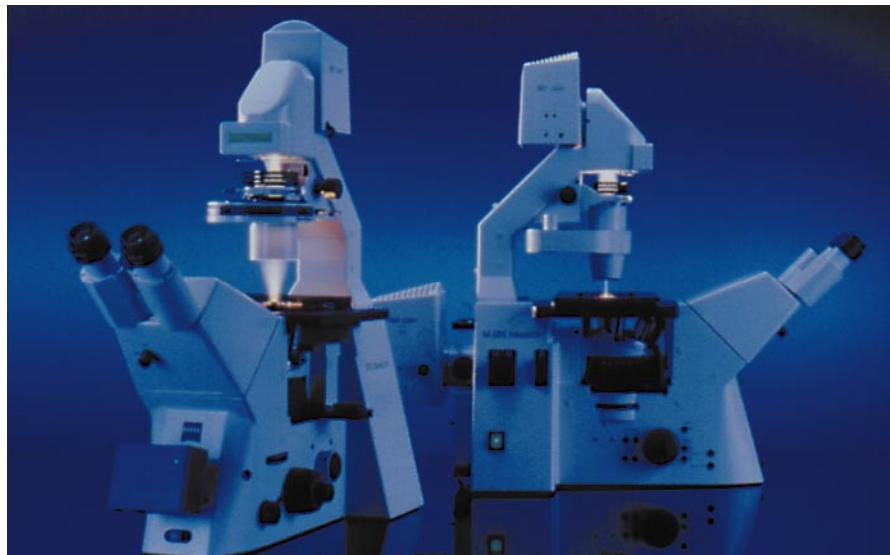
Zeiss

www.zeiss.de

Live cell imaging

The Axiovert 200 inverted microscope has automatic light intensity adjustment, a focus stop to prevent the objective from hitting the stage, a DIC slider that can be operated from both sides, and an Aqua Stop feature that uses silicone rings to protect the objectives, nose-piece and fluorescence filters from contact with liquids. Working distances of up to 70 mm are designed to give a good view and facilitate specimen accessibility when manipulation equipment is being used. Observation and documentation is provided by the Cell Observer software, which permits time-lapse, multichannel, Z-stack, mark and find, and distance measurement applications, and performs angle and statistical calculations.

Reader Service No. 100



Living image: the Zeiss Axiovert 200, running Cell Observer software, is optimized for work on live cells.

Immunostaining

Vector Laboratories

www.vectorlabs.com

Brighten up your walls

A free poster from Vector Laboratories shows immunohistochemically stained tissue sections illustrating the company's peroxidase and alkaline phosphatase substrates in single and multiple labelling; tables on substrate/counterstain compatibility and preferred combinations of substrates for optimal con-

trast in multiple labelling protocols are also included. The poster is free on request from Vector.

Reader Service No. 101

FLIP antibodies

Alexis

www.alexis-corp.com

FLICE-inhibitory protein

An inactive homologue of caspase-8, FLIP is a modulator of apoptosis and survival pathways and is believed to be an antagonist to the death-receptor-initiated apoptosis pathways, which use caspase-8. Improved specificity to FLIP is promised for this new duo of FLIP monoclonal antibodies. The first recognizes human FLIP and is suitable for flow cytometry, while the second recognizes both human and mouse FLIP and is designed for ELISA, immunoprecipitation and western blotting.

Reader Service No. 102

FLIM

La Vision

www.lavision.de

Fluorescence lifetime imaging and microscopy

The excited-state lifetime of a fluorescent molecule is highly sensitive to its molecular state (conformation and orientation) and environmental factors such as pH, ion concentration, viscosity and species. The FLIM system allows investigation of time- and spatially resolved structures, contents and processes on cells by measuring the time-gated fluorescence of pulsed laser-excited molecules. It consists of an excitation laser source, microscope, filter or imaging spectrograph, ICCD camera and accompanying

software. Time resolution is 50ps with a derived spatial resolution in the nm range. The resulting measurements provide information about ligand-receptor interactions, membrane potential, absorption and diffusion kinetics and molecular structure.

Reader Service No. 103

Cytotoxicity fluoro-test

Wako

www.wakousa.com

Cell viability assay

Based on a fluorometric assay of cell lysis and screening, this test utilizes a fluorescent dye to differentiate between intact and damaged cells. The two-step procedure can be used with both adhesive and floating cells. No hazardous wastes are generated.

Reader Service No. 104

CaspaTag

Intergen

www.intergen.com

Fluorescein caspase activity kit

FAM-VAD-FMK, a derivative of the broad-spectrum caspase inhibitor zVAD-FMK, is used to detect active caspases in living cells. The FAM-labelled caspase inhibitor is claimed to produce a brighter signal and more hydrolysis-resistant carboxamides than FITC-labelled peptide conjugates. This kit contains inhibitor, wash, solution, fixative, propidium iodide and Hoechst stain for multi-colour analysis. It is suitable for detection of caspase activation in apoptosis models and studies of the role of interleukin-1- β converting enzyme and its homologues in inflammatory responses.

Reader Service No. 105



Poster session: butterfly yes, ballot no.

new on the market

ImmobaSil

Ashby Scientific www.ashby-scientific.co.uk

Cell support matrix

This system supports the growth of both adherent and suspension cells in culture. Its porous structure provides a large surface area and encourages optimum nutrient flow, allowing high cell densities to be achieved. ImmobaSil is inert and non-toxic, and is bio-compatible with delicate mammalian cell lines. The gas-permeable silicon matrix enhances oxygen transfer and extends the lifespan of long-term cultures. It is resistant to biological and chemical attack and protects cells from shear forces. Typical cell attachment time is two hours, and cell densities in excess of 1×10^8 per ml of carrier can be achieved.

Reader Service No. 106

Retrievit

InnoGenex www.innogenex.com

A fetching formulation

Enhance immunostaining of formalin-fixed tissues in immunohistochemistry and *in situ* hybridization with Retrievit solutions, which optimize signal-to-noise ratios for different antigen-antibody pairs. The solutions are available in formulations suitable for nuclear or cytoplasmic antigens, or those located in the extracellular matrix or membrane surface.

Reader Service No. 107



Allsorts: ImmobaSil comes in all shapes and sizes.

RosetteSep

StemCell Technologies www.stemcell.com

Killer cocktail

This formulation allows direct enrichment of human natural killer cells from whole blood in a single 20-minute labelling procedure performed immediately prior to a standard Ficoll centrifugation. Unwanted cells pellet with red cells, leaving NK cells highly enriched at the Ficoll-plasma interface. RosetteSep cocktails are also available for other lymphocyte subsets.

Reader Service No. 108

FemtoJet

Eppendorf www.eppendorf.com

Microinjector

Suitable for microinjection of foreign mol-

ecules into living cells, both adherent and in suspension, this compact device has a built-in compressor and offers wide pressure ranges, simple operation and precise pressure control. Principal functions are triggered directly on the device or via a hand or optional foot control. Semiautomatic microinjection into adherent cells is possible when FemtoJet is used in combination with the company's micromanipulator.

Reader Service No. 109

Cytofuge 2

StatSpin www.statspin.com

Low-cost personal centrifuge

Cytologists, haematologists and microbiologists are all potential users of this cytocentrifuge, which can produce on-slide cell preparations from a wide range of body fluids in just five minutes. Weighing 2.5 kg, it is claimed to be one of the smallest cytocentrifuges in the world. The rotor holds up to four slides which can be centrifuged at speeds ranging from 600 to 4,400 r.p.m.; programme times between two and 30 minutes can be selected.

Reader Service No. 110

FastImmune CD4

Becton Dickinson www.bdfacs.com

Intracellular cytokine kits and reagents

For performing cytokine detection in antigen-stimulated human whole blood samples, these CD4 intracellular cytokine detection and three-colour cytokine antibody reagents are available as anti-IFN- γ , anti-IL-2 and anti-TNF- α kits. The kits are optimized and pre-mixed to reduce pipetting steps during the assay and remove the need for antibody titrations.

Reader Service No. 111

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

Look no further
for the best value in
Oligo Synthesis

Competitive pricing
Routine 24 hour dispatch • Every Oligo QC verified by PAGE
Rigorous batch QC by PCR and Mass Spec
Web ordering and tracking • Local Freephone numbers
A great deal in any language!!

SIGMA (+44) (0) 1223 833000 • fax (+44) (0) 1223 833200
email: info@sigma-genosys.co.uk • www.sigma-genosys.co.uk

READER ENQUIRY NO. 33

mtDNA EXTRACTOR KITS

NEW

- Easy & efficient isolation of mtDNA from whole blood, cell culture or tissue
- Results in ~90 Minutes
- No Hazardous Reagents
- Works even in the presence of anticoagulants

Wako Wako USA 877-714-1921 Wako GmbH (49)02131 3110
www.wako-chem.co.jp

READER ENQUIRY NO. 38

www.AlphaInnotech.com

New Visions
in Digital Imaging

Alpha Innotech Corporation

READER ENQUIRY NO. 48

Diagnostic Reagents

Pefachrome®-Series: chromogenic and fluorogenic substrates

Pefabloc®-Series: serine proteinase inhibitors

Purified snake venom proteins (e.g. Protac®, Ecarin, Batroxobin)

PENTAPHARM LTD
CH 4002 Basel Fax: +41 61 319 96 19
http://www.pentapharm.ch

READER ENQUIRY NO. 16

BACs • YACs • PACs cDNAs

Research Genetics provides clones and colony membranes for a number of human, mouse and rat libraries as well as a custom screening service for the libraries.

Research Genetics

U.S. or Canada 800-533-4363

U.K. 0-800-89-1393

FAX 256-536-9016

Check out our homepage at
<http://www.resgen.com>

READER ENQUIRY NO. 40

Tackling disease head-on

New US institutes are springing up as the search for a vaccine to combat AIDS intensifies

p753

Emergency aid

A new initiative promises to deliver vaccines to the frequently overlooked developing world

p754

Vaccine centres unite specialists in the battle against infectious diseases

The race to produce an AIDS vaccine is just part of the story at some major new research centres working to develop vaccines, says Diane Gershon.

In May 1997, US President Bill Clinton challenged the scientific community to develop an AIDS vaccine within a decade. This intensified effort included the creation of the Dale and Betty Bumpers Vaccine Research Center at the National Institutes of Health in Bethesda, Maryland.

The new building for the centre is now all but complete and will be officially opened in spring 2001. The five-storey glass building, costing between \$35 million and \$40 million, has around 5,100 square metres of space, four research floors, a Biosafety Level-3 (BL-3) containment facility, conference rooms, a vivarium and a cybercafé. It is open and airy, with lots of common areas to facilitate interaction between staff, who — although drawn from an array of disciplines — are directed towards a common purpose.

"We really want to capture under one roof the different aspects of vaccine science. Our investigators start as basic as X-ray crystallographers, and go as applied as AIDS vaccine clinical trials specialists, and everything in between," says the centre's director Gary Nabel, whose own earlier research interests were in HIV gene therapy and DNA-based therapeutic vaccines for cancer.

Although the primary goal is to develop an AIDS vaccine, he says, "we do have a broader mission, which is to advance vaccines for all diseases". Nabel, for example, continues to work on developing a vaccine against the Ebola virus. As he points out, what is learnt with other diseases could help the research into AIDS, and vice versa. The challenge in the early years, he says, will be in striking the right balance.

Jointly funded by the NIH's National Institute of Allergy and Infectious Diseases and the National Cancer Institute, the centre had a bud-



Work in progress: two new vaccine development institutes are hunting the elusive AIDS vaccine — the Vaccine and Gene Therapy Institute at Oregon Health Sciences University (top) and the National Institutes of Health Vaccine Research Center (left).

get of \$26 million for fiscal year 2000, although "that number needs to rise as we fill the centre", Nabel says. It will focus on the preclinical and early clinical stages of development and will work closely with the newly created HIV Vaccine Trials Network, which will conduct all phases of clinical trials and replaces the AIDS Vaccine Evaluation Group (AVEG) and the HIV Network for Prevention Trials.

Newly appointed clinical director Barney Graham will be the centre's main link to the vaccine trials network, having had more than 10 years of experience with the AVEG. "We will conduct early-stage trials on the NIH campus, mainly so we can learn more scientifically about the immune responses in people, but there will be a passing of the baton

from us to them as we identify our more promising candidates," says Nabel.

The centre will also do research with non-human primates to glean "meaningful information about protection, about the character of the immune responses, and any possible correlates in between," says Nabel. What is learned in humans and primates will then be used to choose vaccine candidates for more advanced clinical testing. The earliest candidates will probably be DNA vaccines, says Nabel.

Mario Roederer, director of the centre's flow cytometry laboratory — and, like almost all the core staff, recruited from outside the NIH — says the new post "gives me the opportunity to bring my technology to a much larger group of scientists, and to apply



Gary Nabel is focusing research on AIDS.

get of \$26 million for fiscal year 2000, although "that number needs to rise as we fill the centre", Nabel says. It will focus on the preclinical and early clinical stages of development and will work closely with the newly created HIV Vaccine Trials Network, which will conduct all phases of clinical trials and replaces the AIDS Vaccine Evaluation Group (AVEG) and the HIV Network for Prevention Trials.

▶ this technology to clinical settings, specifically clinical trials and vaccine trials”.

Even when all the key individuals are in place next spring, there will be spare capacity to allow for future growth and visiting scholars. Although there are no plans at this point to hire any more tenure-track researchers, there will still be a need to recruit staff scientists, postdocs and graduate students.

When asked whether he feels the clock ticking, Nabel says he sees the ten-year time frame more as a point by which the scientific community should be able to say whether a vaccine is a viable approach to containing AIDS. “My guess is that there will be some kind of a vaccine within that time-frame but I don’t know whether it’s going to be the optimal vaccine, because the [clinical] trials take so long.”

Vaccines beside gene therapy

Oregon Health Sciences University (OHSU) — noting that the tools and resources needed for vaccine development have much in common with those of gene therapy — has established a new institute that will span both areas, in a cooperative venture with the Oregon Regional Primate Center.

Although the university had staff with expertise in vaccine development and gene therapy, “the idea was to create a critical mass that would be located in one place”, says Jay Nelson, director of the new Vaccine and Gene Therapy Institute, who has been recruiting heavily over the past year.

In February, the institute will move to a new \$35-million building on OHSU’s new campus in Hillsboro. This will provide 1,800 square metres of laboratory space, 900 square metres of animal facilities, plus four BL-3 and eight BL-2 containment facilities. The institute is next-door to both the primate centre and the Oregon Graduate Institute, which is merging with OHSU. Its strengths in computational science and chemistry complement the functional-genomics aspects of the new institute.

The first phase of recruitment for the new institute has been in vaccine research, with eight of nine core faculty slots already filled, and recruitment continuing at the graduate, postdoctoral and research associate level.

The first new recruit was Louis Picker, director of the institute’s vaccine programme. Picker, who studies T-cell biology in human and non-human primates, will be working on prophylactic and therapeutic vaccines against HIV, its primate equivalent (SIV) and cytomegalovirus. Klaus Früh, director of the institute’s functional-genomics centre, will use approaches such as DNA microarrays, proteomics and laser-capture microscopy, to look at how viruses infect cells.

The next phase of expansion for the institute will be to build up a programme for gene therapy alongside the vaccine programme. ■

Diane Gershon is Assistant Editor, New Technology, Nature Medicine.



JORGEN SCHYTT/STILL

Developing countries and poorest people targeted by UN centre

Almost a third of all deaths worldwide are caused by infectious diseases, with developing countries being the hardest hit. Although vaccines can be effective against such diseases, vaccine development is a lengthy and costly business. To obtain a return on their investment, companies often sell vaccines at prices that are beyond the reach of developing nations. Moreover, the dire economic situation of many of these countries gives companies little incentive to develop vaccines specifically for the developing world.

In October 1997, the United Nations Development Programme, recognizing the need for more research in this area, launched the International Vaccine Institute (IVI) to speed up the introduction of vaccines to developing countries. The Pacific Rim was targeted as a potential location for the new independent, non-profit institute because it had countries with the necessary scientific infrastructure but was also close to poorer countries, says John Clemens, who became IVI director in July 1999. The South Korean government made a successful bid to host the new institute, which is located in temporary accommodation at Seoul National University.

The IVI is holding off on most of its laboratory activities until construction of a new research facility and vaccine pilot-plant is complete in 2002. Consequently, much of the institute’s effort has so far centred on clinical and field activities in the applied vaccine sciences. The initial focus has been on tackling Asian public-health problems, such as enteric diseases such as enterotoxigenic *Escherichia coli*, shigellosis, typhoid fever and cholera; respiratory infections such as *Haemophilus influenzae* type B and pneumococcal disease; and vector-borne diseases such as Japanese encephalitis. But Clemens, who has worked on vaccines for developing countries for 20

years, hopes soon to turn attention to sub-Saharan Africa with the IVI’s Diseases of the Most Impoverished project. This will target cholera, dysentery and typhoid fever with support from a five-year \$40-million grant from the Bill and Melinda Gates Foundation.

IVI staff are currently conducting some basic research into *Shigella* vaccines — “a very appropriate niche for an institute like ours”, says Clemens. *Shigella* attracts very little industrial interest, he says, although it causes more than one million deaths each year in developing countries. The IVI has no plans to develop vaccines commercially, but will strengthen ties with industry and other public-sector organizations to push promising candidates through to clinical practice.

At present the institute has fewer than 40 staff, but eventually Clemens expects to have 200, a quarter of whom will be scientists. He hopes soon to recruit a deputy director of the institute for the laboratory science programme. Most of the scientists recruited to date are biostatisticians, clinical epidemiologists and data managers. As the new laboratory space nears completion, the IVI will also begin recruiting staff with expertise in vaccine production, process research, molecular genetics (particularly microbial genetics), human immunology and microbiology.

Scientific appointments are for fixed terms — something that Clemens says is typical of international research institutes. Clemens himself is on a five-year contract, on leave from the US National Institutes of Health. Jobs in international institutes tend to attract people who “see the uniqueness of the professional opportunity for productivity”, says Clemens. “Most people who come to work at an international institute would not do so with the idea that this would be their life-long career,” he says. **D.G.**